

3 1761 10374544 4













Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761103745444>













Statistics Canada Statistique Canada

Government  
Publications

12-001

---

# **SURVEY METHODOLOGY**

---

**December 1981**

---

**Volume 7**

---

**Number 2**

---

---

A Journal produced by  
Methodology Staff  
Statistics Canada

---

Canada





## SURVEY METHODOLOGY

December 1981

Vol. 7

No. 2

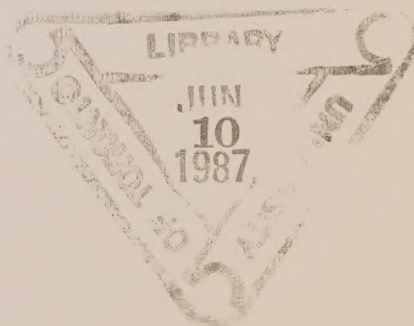
A Journal produced by Methodology Staff, Statistics Canada

### C O N T E N T S

Notes on Inference Based on Data From Complex Sample Designs GAD NATHAN .....	109
The Non-Response Problem J.G. BETHLEHEM and H.M.P. KERSTEN .....	130
On the Variances of Asymptotically Normal Estimators From Complex Surveys DAVID A. BINDER .....	157
An Overview of Canadian Health Statistics: Past, Present and Future LORNE ROWEBOTTOM .....	171
Models for Estimation of Sampling Errors P.D. GHANGURDE .....	177

8-3200-501  
Reference No.  
Z - 079

ISSN: 0714-0045







## SURVEY METHODOLOGY

December 1981

Vol. 7

No. 2

A Journal produced by Methodology Staff, Statistics Canada.

---

Editorial Board:	R. Platek	- Chairman
	M.P. Singh	- Editor
	P.F. Timmons	
	J.H. Gough	- Assistant Editor

---

### Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department.

### Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 6th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.





NOTES ON INFERENCE BASED ON DATA  
FROM COMPLEX SAMPLE DESIGNSGad Nathan<sup>1</sup>

The problems associated with making analytical inferences from data based on complex sample designs are reviewed. A basic issue is the definition of the parameter of interest and whether it is a superpopulation model parameter or a finite population parameter. General methods based on a generalized Wald Statistics and its modification or on modifications of classical test statistics are discussed. More detail is given on specific methods-on linear models and regression and on categorical data analysis.

## 1. INTRODUCTION

Standard methods of inference, such as regression, analysis of variance or tests of independence, are, in general, based on the assumption that the data are obtained by simple random sampling from an infinite population with a probability distribution belonging to some hypothetical family. The wide dissemination of standard computer packages has made the use of these methods extremely easy. However standard methods cannot usually be simply applied to data from complex sample designs without any modification.

In the following we attempt to provide a selection of some practical hints on what can be done and of some warnings against what should not be done in these situations. This is based on the selected list of references to recent work in the area, which include many examples of applications.

The first question which must be answered by anyone who intends to carry out statistical analysis is what exactly are the parameters about which inference is required.

---

<sup>1</sup>G. Nathan, Hebrew University, Jerusalem and Isreal Central Bureau of Statistics



One of two extreme answers to this question is often given (Brewer and Mellor (1973); Smith (1976)). One, as advanced for instance by Kish and Frankel (1974), considers that the only relevant inference concerns finite population parameters, such as the population regression coefficient:

$$B = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2},$$

similarly defined multiple or partial correlation coefficients or other measures, defined with respect to the finite population only, with no recourse to any superpopulation model. Inference in this case would usually be design-based (Sarndal (1978)), that is based only on properties of the sample distribution. However model-based inference about a finite population parameter is also possible (Hartley and Sielken (1975)).

The other extreme position, as stated, for instance, by Fienberg (1980), considers all inference as relating to the parameters of a probability distribution (a superpopulation) of which the finite population represents a realization. Examples of such inference can be found in Konijn (1962), Fuller (1975), Thomsen (1978) and Pfeiffermann and Nathan (1981). If the parameters about which inference is made relate to a superpopulation model, design-based inference cannot be used alone and inference must be model-based, Sarndal (1978), or jointly model- and design-based. Under assumptions of independence between the model distribution and the sampling distribution, standard (model-based) inference is valid and the sample design only affects the efficiency of inference.

Serious objections can be raised with respect to each of these extreme approaches. Model-based inference relies heavily on assumptions about a theoretical model which are usually difficult to ensure and the inference will not, in general, be robust to departures from this model. On the other hand, the finite population parameters, on which design-

based inference is made, are usually "copies" of theoretical model parameters with little descriptive value in themselves, unless some basic model is assumed. For instance, a finite population correlation coefficient is a useful measure of the relationship between two variables only if the relationship is approximately linear.

In many cases some balance between these approaches may be preferable. This can be attained, for instance, by considering as the objects of inference only finite population parameters which closely approximate superpopulation parameters of a suitable model, to which the data fit. For instance, if separate regression equations are fitted to relevant sub-populations a better linear fit may be obtained than from an over-regression. If the sub-populations are large enough this will ensure that the finite population regression coefficients closely approximate the superpopulation parameters, so that any inference relating to the finite population parameters can be considered as relating to the superpopulation parameters.

To ensure close correspondence between model parameters and finite population parameters extensive exploratory analysis to check the model should be carried out, before entering into any formal analysis. This analysis to explore various alternative models can often be based on simple descriptive measures for which the sample design can be taken into account or on graphical displays. However the results have to be carefully interpreted in the light of the sample design. For example, a few large residuals with small sample weights may be much less important than many smaller residuals with large weights. A useful diagnostic tool to consider in the case of regression is the difference between a weighted and an unweighted regression coefficient. A large difference will often indicate that the model is inadequate.

Once the parameters have been determined, we should consider what type of inference is required (point estimation, interval inference or tests of hypotheses). While point estimation and confidence intervals would



be most appropriate for finite population parameters, tests of hypotheses, and in particular simple hypotheses, are strictly relevant only with respect to superpopulation parameters of a well-defined model. For example the hypothesis that two domain means are equal can only be seriously entertained with respect to the superpopulation means rather than their finite population realizations. If one wishes to avoid the formulation of a model it would be preferable to use point estimation or confidence intervals for the difference between the domain means rather than tests of hypotheses. If hypothesis testing about finite population parameters is required, testing a composite hypothesis (e.g. that the difference between the means is in a given range of values) would be more appropriate than testing the simple hypothesis (that the difference is zero). Note that for sufficiently large samples, any non-zero difference, no matter how small, will be found significantly different from zero.

In the following, we discuss some basic general methods of analysis of data from complex sample designs and some specific methods for linear models and for tests of goodness of fit and of independence in contingency tables. In general we shall consider the inference as relating to finite population parameters. However we consider this inference as relevant only if the finite population parameters closely approximate superpopulation model parameters. This leaves open the possibilities of tending either towards a purely design-based approach or towards a purely model-based approach, according to one's personal degree of belief in the validity of an underlying model.

## 2. BASIC GENERAL METHODS

### 2.1 Generalized Wald Statistic

If the hypothesis to be tested is linear (or can be linearized) in the expected values of asymptotically normal statistics, for which a consistent estimator of the variance matrix is available, the generalized Wald Statistic can be used (Grizzle, Starmer and Koch (1969)),

Koch, Freeman and Freeman (1976), Freeman, Freeman, Brock and Koch (1976), Shah, Holt and Folsom (1977) and Koch, Stokes and Brock (1980)).

We assume that we wish to test the hypothesis:

$$H_0: X\beta = \theta_0, \quad (2.1.1)$$

where  $X$  is a known  $r \times p$  design matrix of full rank.  $\beta$  is a  $p \times 1$  unknown parameter vector (either finite population parameters or superpopulation parameters) and  $\theta_0$  is a known  $r \times 1$  vector of constants. In case the hypothesis is not linear a first-order Taylor series approximation can be used (Nathan (1972) and Shuster and Downing (1976)).

We assume that a consistent asymptotically normal estimator  $\hat{\beta}$ , of  $\beta$  is available, as well as a consistent estimator,  $\hat{V}$ , of the covariance matrix of  $\hat{\beta}$ , whose distribution is independent of that of  $\hat{\beta}$ .

Then the generalized Wald Statistic, defined as:

$$X_W^2 = (X\hat{\beta} - \theta_0)' (X\hat{V}X')^{-1} (X\hat{\beta} - \theta_0) \quad (2.1.2)$$

is asymptotically distributed, under the null hypothesis, as chi-square with degrees of freedom equal to the dimension of the hypothesis ( $p-r$ ).

The consistency of  $\hat{\beta}$  and of  $\hat{V}$  and the asymptotic distributions of  $\hat{\beta}$  and of  $X_W^2$  can all be considered with respect to the sampling distribution or with respect to the superpopulation distribution.

The major problem associated with this approach is in obtaining the consistent estimator,  $\hat{V}$ , of the covariance matrix when  $\hat{\beta}$  is non-linear in the sample observations (as will often be the case). Rao (1975) surveys the various methods of variance estimation which can be used: linearization (Tepping (1968)); Balanced Repeated Replication (McCarthy (1969)); and Jackknife (Miller (1974)). Several general computer programmes are available for their implementation - e.g. SUPERCARP (Hidioglou, Fuller and Hickman (1980)), SUDAAN (Shah (1978)) for

linearization and OSIRIS IV: PSALMS for balanced repeated replication. A complete listing and comparison of programs is given by Kaplan, Francis and Sedransk (1979).

Empirical comparisons of the variance estimators are given by Kish and Frankel (1974) and by Richards and Freeman (1980) and theoretical comparisons by Krewski and Rao (1981).

However, attention should be given to the stability of the variance estimator, especially when the number of parameters is large. In addition, care must be taken with respect to the conditions under which consistency and asymptotic properties hold for complex designs. For instance, for a two-stage design asymptotic results may require both a large number of PSU's and a large number of final units per PSU.

## 2.2 Approximation and Modelling of the Covariances

The practical difficulties involved in obtaining a stable consistent estimator of the covariance matrix have led to attempts to use simplified approximations to such estimators. The basic idea is that by assuming some structure for the covariance matrix, more stable estimators of fewer parameters can be used.

The approximation can be carried out under a pure design-based approach, directly with respect to the covariance matrix. If assumptions can be made on equality of design effects for variances and covariances within a given sub-group of parameters, overall estimators of covariance can be used. This approach is used, for instance, by Nathan (1973), Fuller and Rao (1978), Fellegi (1980) and Lepkowski and Landis (1980).

Alternatively modelling of the population structure itself can lead to simplified covariance matrices which can easily be estimated (see, e.g., Altham (1976), Fuller and Battese (1973), Tomberlin (1979), Holt, Richardson and Mitchell (1980), Imrey, Sobel and Francis (1980) and Pfeiffermann and Nathan (1981)).



## 2.3 Modifications of Standard Tests

The widespread use of standard computer packages has encouraged the search for simple modifications to standard test procedures to take into account complex sample design. The idea can be regarded as a natural extension of the use of design effects as multiplicative factors for variances based on a simple random sample of the same size, in order to correct for the complex design used.

The correction may indeed be based on design effects of various estimators or on average design effects (see, e.g., Cowan and Binder (1978), Fay (1979), Fellegi (1980), Rao and Scott (1981) and Scott and Holt (1981)).

Another alternative is to investigate the behaviours of standard test statistics under some superpopulation model and to modify the standard statistic accordingly (Cohen (1976) and Campbell (1977)).

## 3. SPECIFIC METHODS

### 3.1 Linear Models and Regression

The prior determination of the model and of the parameters of interest is extremely important for the case of regression analysis and of linear models. For instance, when different regression relationships must be assumed for different strata or for different PSU's in a two-stage design, the parameter of interest could be a simple average of the regression coefficients (Konijn (1962)); a weighted average of the coefficients (Pfeffermann and Nathan (1981)); or their expected value (under some prior distribution) (Porter (1973)).

The model and the parameters of interest should, in general, be determined on the basis of the assumed overall population structure and should not reflect to the structure of the sample design. However in many cases the sample design will reflect population structure so that

sample design variables may be part of the model. For example consider the model:

$$E(Y|X_1, X_2) = X_1 \beta_{1.2} + X_2 \beta_{2.1} \quad (3.1.1)$$

where  $X_1$  includes only variables which do not relate to the sample design and  $X_2$  includes all the variables which enter into the complex sample design, i.e. the sample distribution depends only on  $X_2$ :

$$P(s|X_1, X_2) = P(s|X_2). \quad (3.1.2)$$

The estimation of  $\beta_{1.2}$  and of  $\beta_{2.1}$  in (3.1.1) and inference about them can proceed in the classical way, as if sampling were simple random, if indeed (3.1.1) holds.

However if the design variables,  $X_2$ , are not included in the regression equation of interest:

$$E(Y|X_1) = X_1 \beta_1 \quad (3.1.3)$$

and the design variable  $X_2$  is correlated with  $Y$  (conditional on  $X_1$ ) then the standard OLS estimator of  $\beta_1$  is not consistent (see Nathan and Holt (1980) and Holt and Smith (1979), who propose modified weighted and unweighted estimates of  $\beta_1$ , which are consistent). Holt, Smith and Winter (1980) give an example of the application of these estimators.

If the linear model:

$$E(Y_i|x_i) = x_i' \beta \quad (3.1.4)$$

$$\text{cov}(Y_i, Y_j | x_i, x_j) = \begin{cases} \sigma^2 & i=j \\ 0 & i \neq j \end{cases} \quad (3.1.5)$$

indeed holds for all population units ( $i, j=1, \dots, N$ ) of a finite population and the  $p \times 1$  column vector  $x_i$  includes all the sample design variables, then the OLS unweighted estimator:

$$\hat{\beta} = (X_n' X_n)^{-1} X_n' Y_n \quad (3.1.6)$$

based on the sampled values  $X_n' = (x_1, \dots, x_n)$  and  $Y_n' = (y_1, \dots, y_n)$

is the "best" linear model-unbiased estimator of  $\beta$  irrespective of the sample design. "Best" here is in the sense of minimal model-variance. However  $\hat{\beta}$  is, in general, not a design-unbiased, nor even a design-consistent, estimator of the population parameter:

$$B = (X_N' X_N)^{-1} X_N' Y_N, \quad (3.1.7)$$

where  $X_N' = (x_1, \dots, x_N)$  and  $Y_N' = (y_1, \dots, y_N)$ .

The design-consistent estimator of  $B$  is the weighted estimator:

$$\hat{\beta}_W = (X_n' W_n X_n)^{-1} X_n' W_n Y_n, \quad (3.1.8)$$

where the weight matrix,  $W_n = \text{diag} (\pi_1^{-1}, \dots, \pi_n^{-1})$ , is the  $n \times n$  diagonal matrix of the reciprocals of the sample inclusion probabilities  $\pi_i = \Pr(i \in s)$ .

The consistency of  $\hat{\beta}_W$ , as an estimator of  $B$ , obviously does not depend on the model (3.1.4) holding, but the relevance of estimating  $B$  when the model does not hold can be challenged. It can be shown that under certain conditions for a non-linear model, which assumes that the conditional expectation of  $Y$  (given  $X$ ) is a differentiable function of  $X$ , the model-expectation of  $B$  can be expressed approximately as a weighted average of the slopes of this function at the points  $X_i$  (the weights depending only on  $X_i - \bar{X}$ ). However this interpretation is of limited practical value.

In any case  $\hat{\beta}_W$  is a model-unbiased estimator of  $\beta$ , whenever (3.1.4) does hold. It will not, in general, be an optimal estimator of  $\beta$  under (3.1.5) for unequal probability sampling, but will be so if the conditional model variance of  $Y_i$  is proportional to  $\pi_i$ ,



i.e. 
$$V(Y_i | x_i) = k \pi_i . \quad (3.1.9)$$

Since the weighted estimator,  $\hat{\beta}_W$ , is more robust than the unweighted estimator,  $\hat{\beta}$ , in the sense that it is both a model-unbiased estimator of  $\beta$ , if the model holds and a design-consistent estimator of  $B$ , if not, the use of the weighted estimator  $\hat{\beta}_W$  is recommended, for estimation of  $B$ , whenever there is no assurance that the model (3.1.4) - (3.1.5) holds. The question which must then be answered by the subject-matter specialist is whether  $B$  is a relevant parameter to estimate.

It should be noted that for self-weighting designs  $\hat{\beta}$  and  $\hat{\beta}_W$  coincide. The estimator,  $\hat{\beta}_W$  (3.1.8), can be obtained directly from standard computer programmes which provide for weighted regression (e.g. BMDP) by using the weights  $1/\pi_i$ ; or from other programmes (e.g. SPSS) by carrying out unweighted regression on the transformed variables  $Y_i/\sqrt{\pi_i}$  and  $x_i/\sqrt{\pi_i}$ , but not on the weighted variables  $Y_i/\pi_i$ ,  $x_i/\pi_i$ . However, it should be noted that under either alternative the reported variances and covariances of the estimators are incorrect and that the standard significance tests (e.g. F tests) are invalid, and can result in grossly misleading conclusions.

Assuming the model (3.1.4) - (3.1.5), the model variance of  $\hat{\beta}$  is:

$$V(\hat{\beta} | X_n) = \sigma^2 (X_n' X_n)^{-1} , \quad (3.1.10)$$

which is the result given by standard unweighted regression programmes. However, the model variance of  $\hat{\beta}_W$  is:

$$V(\hat{\beta}_W | X_n) = \sigma^2 (X_n' W_n X_n)^{-1} X_n' W_n' W_n X_n (X_n' W_n X_n)^{-1} . \quad (3.1.11)$$

The weighted regression programme, with weights  $1/\pi_i$ , will give a value of  $(X_n' W_n X_n)^{-1}$  for the model variance of  $\hat{\beta}_W$ , which equals (3.1.11) only if  $W_n = I_n$ . Thus none of the standard outputs for standard errors or for tests of hypotheses are correct.

However the estimator of the multiple correlation coefficient obtained from weighted regression:

$$\hat{R}^2 = \frac{(Y_n - X_n \hat{\beta}_W)' W_n (Y_n - X_n \hat{\beta}_W)}{(Y_n - \bar{y}_n \mathbf{1}_n)' W_n (Y_n - \bar{y}_n \mathbf{1}_n)}, \quad (3.1.12)$$

where  $\bar{y}_n = (\sum_s Y_i / \Pi_i) / (\sum_s 1 / \Pi_i)$ , is a design-consistent estimator of the population multiple correlation coefficient:

$$R^2 = \frac{(Y_N - X_N B)' (Y_N - X_N B)}{(Y_N - \bar{Y}_N \mathbf{1}_N)' (Y_N - \bar{Y}_N \mathbf{1}_N)} \quad (3.1.13)$$

where  $\bar{Y}_N = (1/N) \mathbf{1}_N' Y_N$ .

The design-variance of  $\hat{\beta}_W$ , which must be considered the relevant measure of accuracy for  $\hat{\beta}_W$  as an estimator of  $B$ , cannot in general, be obtained from only the first order inclusion probabilities,  $\Pi_i$ . For most sample designs used in practice, the design-variance of  $\hat{\beta}_W$  will have to be estimated by one of the variance estimating techniques mentioned above i.e. linearization, Balanced Repeated Replication or Jackknife (see, e.g., Jonrup and Remmermalm (1976) and Holt and Scott (1981)).

### 3.2 Categorical Data Analysis

The simplest analysis of categorical data relates to a single classification of the population into  $k$  classes with probabilities (relative frequencies)  $\underline{p}' = (p_1, \dots, p_{k-1})$ . In order to test the null hypothesis of goodness of fit to a known distribution  $\underline{p}_0' = (p_{01}, \dots, p_{0k-1})$ :

$$H_0: \underline{p} = \underline{p}_0, \quad (3.2.1)$$

the approaches outlined in section two can be used.

We assume that a consistent survey estimator  $\hat{\underline{p}}' = (\hat{p}_1, \dots, \hat{p}_{k-1})$  of  $\underline{p}'$  is available. If it is asymptotically normal:

$$\sqrt{n} (\hat{p} - p) \rightarrow N(0, V) \quad (3.2.2)$$

and a consistent estimator,  $\hat{V}$ , of  $V$  is available, then the generalized Wald statistic:

$$X_W^2 = n(\hat{p} - p_0)' \hat{V}^{-1} (\hat{p} - p_0), \quad (3.2.3)$$

which is distributed asymptotically as  $\chi^2_{k-1}$  under  $H_0$ , can be used to test  $H_0$ .

For many simple designs consistent estimators of  $V$  are directly available and for more complex designs they can be obtained by standard methods. However if tests of hypotheses of goodness of fit have to be carried out for a variety of variables and classifications, the use of the standard  $\chi^2$  statistic:

$$X^2 = n \sum_{i=1}^k (\hat{p}_i - p_{0i})^2 / p_{0i} = n(\hat{p} - p_0)' P_0^{-1} (\hat{p} - p_0), \quad (3.2.4)$$

where  $P_0 = \text{diag}(p_0) - p_0 p_0'$ , with appropriate modification may be preferred. Rao and Scott (1981) show that the asymptotic distribution of  $X^2$  under  $H_0$  is that of a weighted sum of  $k-1$  independent  $\chi^2$  variables with one degree of freedom each.

$$X^2 \rightarrow \sum_{i=1}^{k-1} \lambda_i Z_i^2; \quad Z_i \sim N(0,1) \text{ independent} \quad (3.2.5)$$

where  $\lambda_1, \dots, \lambda_{k-1}$  are the eigenvalues of

$$D = P_0^{-1} V \quad (\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{k-1} > 0). \quad (3.2.6)$$

A conservative test of (3.2.1) can then be obtained by using the statistic  $X^2 / \lambda_1$  in conjunction with a  $\chi^2_{k-1}$  distribution.  $\lambda_1$  can be components of  $\hat{p}$ . For example, for proportional stratified sampling  $\lambda_1 \leq 1$ , so that  $X^2$  itself can be used as a conservative test statistic.

In other cases the use of  $X^2 / \bar{\lambda}$  with:



$$\bar{\lambda} = \frac{1}{k-1} \sum_{i=1}^{k-1} \lambda_i = \frac{1}{k-1} \sum_{i=1}^k d_i (1 - p_i) ,$$

where  $d_i = V[\hat{p}_i]/[p_i(1-p_i)]$  is the design effect for  $\hat{p}_i$ , has been shown to be a good approximative test by Hidioglou and Rao (1981) for the Canada Health Surveys and by Holt, Scott and Ewings (1980) for large scale U.K. surveys. An alternative approximation -  $X^2/\bar{d}$ , where  $\bar{d} = k^{-1} \sum_{i=1}^k d_i$  - has been proposed by Fellegi (1980).

Direct modelling for  $p$  has been proposed by Altham (1976) and by Cohen (1976), but their models have the serious limitation that they imply  $\lambda_1 = \lambda_2 = \dots = \lambda_{k-1} = \bar{\lambda}$ , which is equivalent to a constant design effect over categories. This is not a realistic assumption, in general, and results in  $X^2/\bar{\lambda}$  having exactly an asymptotic  $\chi^2_{k-1}$  distribution.

For testing independence in a two-way contingency table, the hypotheses can be formulated:

$$H_0: h_{ij}(p) = p_{ij} - p_{i+} p_{+j} = 0$$

$$(i=1, \dots, r-1; j=1, \dots, c-1), \quad (3.2.7)$$

where  $p_{ij}$  is the population probability of cell  $(i,j)$   $p_{i+}$ ,  $p_{+j}$  are the marginal probabilities and  $p' = (p_{11}, \dots, p_{rc-1})$ . The generalized Wald statistic for testing  $H_0$  is:

$$X_{WI}^2 = n[h(\hat{p})]' \hat{V}_h^{-1} h(\hat{p}) , \quad (3.2.8)$$

where  $[h(\hat{p})]' = [h_{11}(\hat{p}), \dots, h_{r-1, c-1}(\hat{p})]$  and  $\hat{V}_h/n$  is a consistent estimator of the covariance matrix of  $h(\hat{p})$ . Versions of (3.2.8) for specific designs with various methods for estimating  $\hat{V}_h/n$  have been used by Garza-Hernandez and McCarthy (1962), Nathan (1969, 1975) Shuster and Downing (1976) and Fellegi (1980).

A modified statistic similar to  $\chi^2/\bar{\chi}$  has been proposed by Rao and Scott (1981):

$$\chi_{CI}^2 = (n/\hat{\delta}) \sum_{i=1}^r \sum_{j=1}^c (\hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j})^2 / (\hat{p}_{i+} \hat{p}_{+j}), \quad (3.2.9)$$

where  $\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^r \sum_{j=1}^c \hat{v}_{ij}(\underline{h}) / (\hat{p}_{i+} \hat{p}_{+j})$  and

$\hat{v}_{ij}(\underline{h})/n$  is an estimator of the variance of  $h_{ij}(\underline{p})$ .  $\hat{\delta}$  can be written in terms of the estimated deffs of  $h_{ij}(\underline{p})$ :

$$\hat{\delta} = \frac{1}{(r-1)(c-1)} \sum_{i=1}^r \sum_{j=1}^c (1 - \hat{p}_{i+})(1 - \hat{p}_{+j}) \hat{\delta}_{ij}, \quad (3.2.10)$$

where  $\hat{\delta}_{ij}$  is an estimator of the deff,  $\delta_{ij}$ , of  $h_{ij}(\underline{p})$ :

$$\delta_{ij} = nV[h_{ij}(\underline{p})] / [p_{i+} p_{+j} (1 - p_{i+})(1 - p_{+j})]. \quad (3.2.11)$$

Estimates of the design effects may be easier to obtain than estimates of variances.

Empirical investigations by Holt, Scott and Ewings (1980) and by Hidioglou and Rao (1981) indicate that the distribution of  $\chi_{CI}^2$  is close to  $\chi_{(r-1)(c-1)}^2$ .

### 3.3 Other Types of Analysis

While linear models, tests of goodness of fit and tests of independence cover many important analysis applications, other types of analysis, such as principal component and factor analysis, discriminant analysis, path analysis, logistic regression, log-linear models non-parametric methods, etc. cannot be directly dealt with in the same way. While the general techniques outlined in section two could be

used, their application presents difficulties and only few cases of their application have been reported.

Since correlation coefficients are a basic element in most multivariate analysis, some empirical studies of the effect of sample design on their estimation have been carried out by Kish and Frankel (1974), Bebbington and Smith (1977) and Holt, Richardson and Mitchell (1980). No general conclusions can be formulated, but design effects are definitely not negligible. Bebbington and Smith (1977) have also studied the sampling variability of principal components estimators.

In other areas design effects for logits have been studied by Lepkowski and Landis (1980) and confidence intervals for quantiles by Woodruff (1952) and by Sedransk and Meyer (1978).

#### ACKNOWLEDGEMENTS

This paper has benefited from comments by and discussion with D. Binder, N. Chinnappa, S.E. Fienberg, M. Hidirolou, G.J.C. Hole, J.N.K. Rao and A. Scott.

#### REFERENCES

- [1] Altham, P.M.E. (1976), "Discrete Variable Analysis for Individuals Grouped Into Families", *Biometrika*, 63, 263-269.
- [2] Brewer, K.R. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys", *Aust. J. Statist.*, 15, 145-152.
- [3] Bebbington, A.C. and Smith, T.M.F. (1977), "The Effect of Survey Design on Multivariate Analysis", *The Analysis of Survey Data* (C.A. O'MUIRCHEARTAIGH and C. PAYNE, EDITORS). Vol. 2, Model Fitting, New York: Wiley, 175-192.



- [4] Campbell, C. (1977), "Properties of Ordinary and Weighted Least Squares Estimators for Two Stage Samples", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 800-805.
- [5] Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Clustered Sampling", J. Amer. Statist. Assoc. 71, 665-670.
- [6] Cowan, J. and Binder, D.A. (1978). "The Effect of a Two-Stage Sample Design on Tests of Independence", Survey Methodology, Vol. 4, No. 1, 16-29.
- [7] Fay, R.E. (1979), "On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling", Proc. Soc. Statist. Sect., Amer. Statist. Assoc. 402-405.
- [8] Fellegi, I.P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", J. Amer. Statist. Assoc. 75, 261-268.
- [9] Fienberg, S.E. (1980), "The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey", The Statistician, 29, 313-350.
- [10] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankhya C, 37, 117-132.
- [11] Fuller, W.A. and Battese, G.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structure", J. Amer. Statist. Assoc. 68, 626-632.
- [12] Fuller, W.A. and Rao, J.N.K. (1978), "Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix", Ann. Statist. 6. 1149-1158.
- [13] Freeman, D.H. Jr., Freeman, J., Brock, D.B. and Koch, G.G., "Strategies in the Multivariate Analysis of Data from Complex Surveys 11: An Application to the United States National Health Interview Survey", Inter. Statist. Rev. 44, 317-330.

- [14] Garza-Hernandez, T. and McCarthy, P.J. (1962), "A Test of Homogeneity for a Stratified Sample", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 200-202.
- [15] Grizzle, J.E., Starmer, C.F. and Kock, G.G. (1969), "Analysis of Categorical Data by Linear Models", Biometrics, 25, 489-504.
- [16] Hartley, H.O. and Sielken, R.L. (1975), "A Superpopulation Viewpoint for Finite Population Sampling", Biometrics, 31, 411-422.
- [17] Hidioglou, M.A., Fuller, W.A. and Hickman, R.D. (1980). Super Carp: Sixth Edition, Statistical Laboratory Survey Section, Iowa State University, Ames, Iowa.
- [18] Hidioglou, M.A. and Rao, J.N.K. (1981), "Chisquare Tests for the Analysis of Categorical Data from the Canada Health Survey", Invited Paper for 43rd Session of I.S.I., Buenos-Aires.
- [19] Holt, D., Richardson, S.C. and Mitchell, P.W. (1980), "The Analysis of Correlations in Complex Survey Data", (unpublished).
- [20] Holt, D. and Scott, A.J. (1981), "Regression Analysis using Survey Data", The Statistician, 30. (to appear).
- [21] Holt, D., Scott, A.J., and Ewings, P.O. (1980), "Chi-Squared Tests with Survey Data", J. Roy. Statist. Soc. A., 143, 302-330.
- [22] Holt, D., Smith, T.M.F. (1979), "Regression Analysis of Data from Complex Surveys", Roy. Statist. Soc. Conf., Oxford.
- [23] Holt, D., Smith, T.M.F. and Winter, P.O. (1980), "Regression Analysis of Data from Complex Surveys", Jour. Roy. Statist. Soc. A, 143, 474-483.
- [24] Imvrey, P., Sobel, E. and Francis, M. (1980), "Modeling Contingency Tables from Complex Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 213-217.

- [25] Jonrup, H. and Rennermalm, B. (1976), "Regression Analysis in Samples from Finite Population", Scand. Jour. Statist., 3, 33-37.
- [26] Kaplan, B., Francis, I., and Sedransk, J. (1979), "A Comparison of Methods and Programs for Computing Variances of Estimators from Complex Sample Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 97-100.
- [27] Kish, Leslie and Frankel, M.R. (1970), "Balanced Repeated Replication for Standard Errors", J. Amer. Statist. Assoc., 65, 1071-1094.
- [28] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples (with discussion)", J. Roy. Statist. Soc. B, 36, 1-37.
- [29] Koch, G.G., Freeman, D.H., Jr., and Freeman, J.L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys", Inter. Statist. Rev. 43, 59-78.
- [30] Koch, G.G., Stokes, M.E. and Brock, D. (1980), "Applications of Weighted Least Squares Methods for Fitting Variational Models to Health Survey Data", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 218-223.
- [31] Konijn, H.S. (1962), "Regression Analysis for Sample Surveys", J. Amer. Statist. Assoc. 57, 590-606.
- [32] Krewski, D., and Rao, J.N.K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods", Ann. Statist., 9 (5) 1010-1019.
- [33] Lepkowski, J.N. and Landis, J.R. (1980), "Design Effects for Linear Contrasts of Proportions and Logits", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 224-229.
- [34] McCarthy, P.J. (1969). "PSEUDO-REPLICATION: Half-Samples", Inter Statist. Rev. 37, 239-264.



- [35] Miller, R.G. (1974), "The JACKKNIFE- A Review", *Biometrika* 61, 1-15.
- [36] Nathan, G. (1969), "Tests of Independence in Contingency Tables from Stratified Samples", *New developments in Survey Sampling* (N.L. Johnson and H. Smith, eds.). New York: Wiley, 578-600.
- [37] Nathan, G. (1972), "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", *J. Amer. Statist. Assoc.*, 67, 917-920.
- [38] Nathan, G. (1973), "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", *National Center for Health Statistics, Vital and Health Statistics Series 2, No. 53*, Washington, D.C.
- [39] Nathan, G. (1975), "Tests of Independence in Contingency Tables from Stratified Proportional Samples", *Sankhya C*, 37, 77-87.  
[corrigendum: *Sankhya C*, 40, (1978), 190].
- [40] Nathan, G. and Holt, D. (1980), "The Effect of Survey Design on Regression Analysis", *J. Roy. Statist. Soc. B*, 42, 377-386.
- [41] Pfeiffermann, D., and Nathan, G. (1981), "Regression Analysis of Data from Complex Samples", *J. Amer. Statist. Assoc.*, 76, 681-689.
- [42] Porter, R.M. (1973), "On the Use of Survey Sample Weights in the Linear Model", *Annals of Economic and Social Measurement*, 2, 141-158.
- [43] Rao, J.N.K. (1975), "Analytic Studies of Sample Survey Data", *Survey Methodology*, Vol. 1, Supplementary Issue.
- [44] Rao, J.N.K. and Scott, A.J. (1981), "The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", *J. Amer. Statist. Assoc.* 76, 221-230.

- [45] Richards, V. and Freeman, D.H. Jr. (1980), "A Comparison of Replicated and Pseudo-Replicated Covariance Matrix Estimators for the Analysis of Contingency Tables", Proc. Sec. Survey Meth., Amer. Statist. Assoc., 209-211.
- [46] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.
- [47] Sedransk, S. and Meyer, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling", J. Roy. Statist. Soc. B., 40, 239-252.
- [48] Scott, A. and Holt, D. (1981), "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods", (unpublished)
- [49] Shah, B.V. (1978), "SUDAAN: Survey Data Analysis Software", Proc. Statist. Comp. Sect., Amer. Statist. Assoc., 146-151.
- [50] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977), "Inference about Regression Model from Sample Survey Data", Bull. Inter. Statist. Inst. 47, Bk. 3, 43-57.
- [51] Shuster, J.J. and Downing, D.J. (1976), "Two-Way Contingency Tables for Complex Sampling Schemes", Biometrika 63, 271-278.
- [52] Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review (with discussion)", J. Roy. Statist. Soc. A., 139, 183-195.
- [53] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 66, 411-414.
- [54] Thomsen, I. (1978), "Design and Estimation Problems when Estimating a Regression Coefficient from Survey Data", Metrika 25, 27-35.

- [55] Tomberlin, T.J. (1979), "The Analysis of Contingency Tables of Data from Complex Samples", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 152-157.
  
- [56] Woodruff, Ralph S. (1952), "Confidence Intervals for Medians and Other Position Measures", J. Amer. Statist. Assoc. 47, 635-646.



## THE NONRESPONSE PROBLEM

J.G. BETHLEHEM AND H.M.P. KERSTEN<sup>1</sup>

This paper presents an outline of the nonresponse research which is carried out at the Netherlands Central Bureau of Statistics. The phenomenon of nonresponse is put into a general frame-work. The extent of nonresponse is indicated with figures from a number of CBS-surveys. The use of auxiliary variables is discussed as a means for obtaining information about nonrespondents. These variables can be used either to characterize nonrespondents or as stratification variables in adjustment procedures.

Adjustment for nonresponse bias by means of subgroup weighting is considered in more detail. Finally, the last section lists a number of other methods which also aim at reduction of the bias.

## 1. INTRODUCTION

Nonresponse is becoming a growing concern in survey research. The phenomenon of nonresponse, when people are not able or willing to answer questions asked by the interviewer, can appear in sample surveys as well as in censuses. It affects the quality of the survey in two ways: first of all, due to reduction of the available amount of data, estimates of population parameters will be less precise. Secondly, if a relationship exists between the variable under investigation and response behaviour, statements made on the basis of the response are not valid for the total population. For example if the housing demand of respondents is greater than the housing demand of nonrespondents, estimates of the housing demand in the total population will be significantly too high.

---

<sup>1</sup> J.G. Bethlehem and H.M.P. Kersten, Netherlands Central Bureau of Statistics. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

It is obvious that the extent of the nonresponse must be kept as small as possible. If, in spite of these efforts, there still remains a considerable amount of nonresponse, measures have to be taken in order to prevent formulation of wrong statements about the population. Combination of adjustment procedures and usual estimation techniques is necessary to yield valid population estimates.

Two departments of the CBS (Netherlands Central Bureau of Statistics) are involved in nonresponse research. The Department for Social Surveys is responsible for the field work of the surveys. It is concerned with minimizing nonresponse during the process of collecting data. Research is carried out on the optimal number of recalls and the time of the interview. (See Widdershoven & Van den Berg (1980).) Experiments are set up to find the optimal way to approach persons and households with introductory letters. Attempts are made to measure the impact of interview fatigue and interview pressure. Ultimately, notwithstanding these efforts, there still remains an amount of nonresponse. The Department for Statistical Methods investigates the effect of nonresponse on the accuracy of the results of the survey. Methods are developed there to adjust population estimates for the bias due to nonresponse. The remainder of this paper is mainly concerned with the work of the latter department.

The next sections present an outline of the nonresponse analysis at the CBS. Section 2 introduces definitions and the accompanying problems. Nonresponse figures of a number of CBS-surveys are summarized. In section 3 graphical methods are discussed to select auxiliary variables. They provide insight into nonresponse and can be used in adjustment procedures. Section 4 presents adjustment methods which make use of subgroup weighting and section 5 lists a number of other methods.

## 2. THE PHENOMENON OF NONRESPONSE

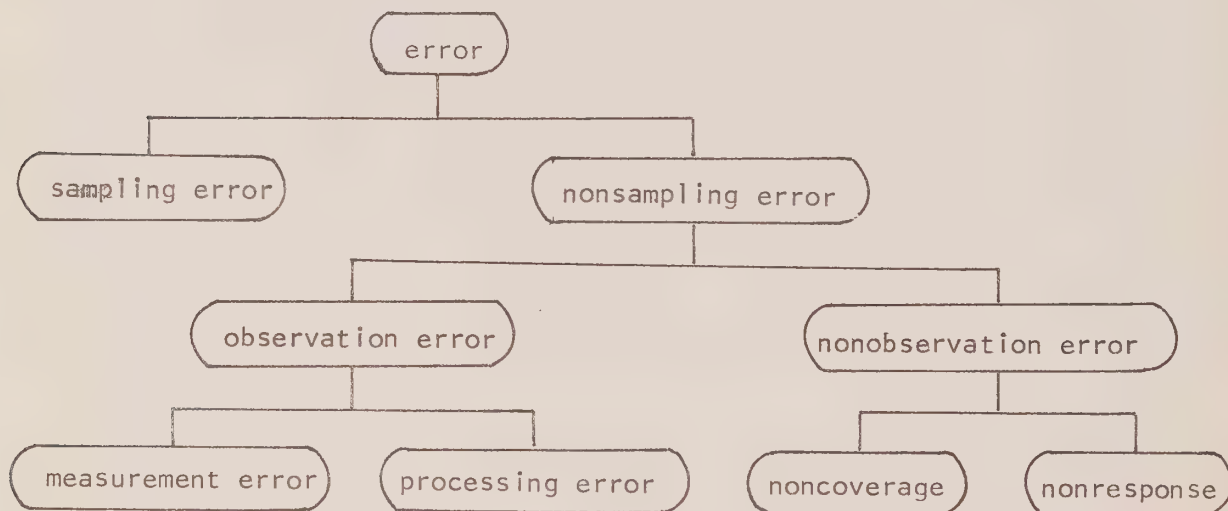
In this section the problem of nonresponse is placed in a general framework, in which also a number of other sampling problems play a role.

Nonresponse figures for a number of CBS surveys are given. Situations are described in which a relationship exists between the variable under investigation and the response behaviour. In the last part of the section two models for the general of nonresponse are considered.

## 2.1 Terminology

The objective of every survey is the determination of certain population characteristics. Due to all kinds of errors, the true value will generally never be obtained. A typology of sources of error is presented in fig. 1. The scheme is due to Kish (1967).

FIG. 1. TYPOLOGY OF ERRORS IN SURVEYS



The two sources of error in surveys are sampling errors and nonsampling errors. Sampling errors consist of that part of the error which is due to the fact that only a sample of values is observed rather than the total population. The sampling error has an expected frequency distribution generated by the totality of sampling errors in all possible samples of the same size. This distribution is used to estimate the population characteristic.

Nonsampling errors are those errors in sample estimates which can not be attributed to sampling fluctuations. Nonsampling errors are often a more serious problem than sampling errors. Nonsampling errors can be divided in observation errors and nonobservation errors.

Observation errors are caused by obtaining and recording observations incorrectly. They may be further subdivided into measurement errors and processing errors.

Measurement errors are caused either by the interviewer or by the respondent. The interviewer himself can be a source of error. He can influence the response by his mere presence, by his (or her) sex, skin colour, age, or dress. Also the way in which he asks questions and clarifies statements affects results. The answer of a person may depend on the type of question (whether a question measures a fact such as year of birth, or an opinion). Errors can also be introduced by factors such as whether the person understands the question, whether he knows the answer or not, whether he wishes to conceal the answer, or whether he wishes to present a certain image. Moreover, memory is not always free of errors, and data may be incorrectly recorded.

Processing errors arise during the processing of the data at the office. They occur during the stage of coding, tabulating and computing.

Nonobservation errors are due to the failure to obtain observations on certain parts of the population. They may be subdivided in noncoverage and nonresponse.

Let the target population be the population the survey is intended to cover. Practical difficulties in handling parts of the population may result in their elimination from the scope of the survey. It is also possible that the actually sampled population contains elements which do not belong to the scope of the survey.



Noncoverage refers to all errors which result from differences between target population and sampled population. Elements which belong to the target population as well as to the sampled population are correct elements. The situation in which elements in the target population do not appear in the sampled population is called undercoverage. These elements have zero probability of selection in the sample. The situation in which elements in the sampled population do not appear in the target population is called overcoverage. Elements, classified as overcoverage, are called duds. They have to be excluded from the sample before analysis takes place. If there is unexpected overcoverage the ultimate sample size may be less than the planned sample size.

Nonresponse refers to failure to obtain observations on some elements selected and designated for the sample. A good classification of nonresponse errors depends on the survey situation. The classification given below focuses on problems in face-to-face interviews. A similar treatment may be applicable in other survey situations. The following categories of nonresponse can be distinguished:

- (1) Not at home. To reduce the extent of this category recalls can be made. Research should be carried out on the optimal number of recalls. The term temporarily unavailable would be a useful generalization for this category, denoting a delay rather than a denial of the interview. The respondent may be too busy, tired, or ill at the time, but will be cooperative on another call.
- (2) Refusal. Some of the factors causing refusal are temporary and changeable. A person may refuse because he is ill-disposed or approached at the wrong hour. Another try, or another approach may find him cooperative. Since quite a number of refusals can, however be considered permanent, a better term for this category is unobtainable,, denoting a denial rather than a delay of observation. Repeated attempts will not bring success. From this view, respondents known to be away during the entire survey period belong in this category, rather than among the not-at-homes.

- (3) Incapacity or inability. This type of nonresponse may refer to mental or physical illness which prevents response during the entire survey period. A language barrier belongs also to this category. If generalized this category could fit in the previously defined unobtainables. It can, however, be useful in some situations to distinguish between the unwilling and the willing, but incapable, respondent.
- (4) Not found. This category can e.g. be large for movers. Such respondents are either not identified or followed because this would be too expensive. Cases of not attempted interviews belong to the same general category. They could be caused by inaccessibility (lighthouse keeper, shepherd), or dangerous surroundings (watchdog, slum).
- (5) Lost information. Information may get lost after a field attempt. Some questionnaires may be unusable because of poor quality or cheating. Other may remain unfilled because they were lost or forgotten.

The typology as described above is applicable in most survey situations, but care must be taken in case of complex sampling designs. When e.g. sampling takes place in more stages the typology can be used in each separate stage. The same source of error can be classified differently in different stage. This is illustrated in an example. In a household survey first a sample of households is selected. The interviewer enumerates all persons in a particular selected household and after that selects a sample from this list. In such an enumeration the student living in an attic is often concealed. In the first stage of the sampling procedure this situation would be classified as measurement error, and in the second stage as undercoverage.

For some sources of error classification may depend on other factors and appropriate rules to cover them must be adopted. For example, if a person to be interviewed died before the interview could take place, classification

depends on the time of death. If death occurred before the day the sample was selected this could be classified as overcoverage, but if death occurred between the day the sample was selected and the day of the interview, the correct classification may be nonresponse.

Before selecting the sample, the population must be divided into sampling units. To every element in the population there must correspond one and only one sampling unit. The construction of the physical list of sampling units, called the sampling frame, is often a major practical problem. The nature of the available sampling frames is an important consideration in sample design. Relevant factors include the type of sampling unit, extent of coverage, accuracy and completeness of the list, and the amount and quality of auxiliary information in the list.

For sampling frames in which the sampling unit is a person the CBS has to restrict itself to administrative records of local authorities (municipalities). For household surveys the CBS manages its own frame, but at the moment the use of the list of delivery points of the Post Office is considered as a sampling frame.

## 2.2 The Extent of Nonresponse

It is rather difficult to compare nonresponse figures of different surveys. The percentage of nonresponse depends on a number of circumstances: aim of the survey, type of sampling unit, the sampling design, efficiency of the field work, performance of the interviewers, nonresponse reducing measures, period in which the survey is held, the target population, the length of the questionnaire, wording of questions, etc. Even the definition of nonresponse may differ. It is necessary to create a frame-work which enables proper comparison of surveys. By controlling the factors which influence nonresponse figures, judgement can be passed on the quality of the different surveys. Such a frame work also offers opportunities for comparing surveys from different countries.

Table 1 presents nonresponse figures of a number of CBS-surveys. A clear trend of increasing nonresponse percentages can be seen in this table.

Table 1: Nonresponse percentages of some CBS-surveys

year	LFS		SSC		SLC		NTS		HS	
	tn	rn	tn	rn	tn	rn	tn	rn	tn	rn
1973	13.2									
1974					28.2	15.6				
1975	15.8	9.0	30.1	18.3					14.5	
1976			28.1	18.6	23.0 <sup>1)</sup>	15.6			12.9	
1977	13.1	6.6	30.9	20.5	29.7	16.9			17.6	9.3
1978			36.1	23.9			33.0	26.2	21.9	12.5
1979	19.7		36.6	24.4	33.7 <sup>2)</sup>		30.6	23.9	25.5	
1980			36.8	24.7	35.6	19.7	32.1	24.5		

1) = elderly people only

LFS = Labour Force Survey

2) = young people only

SSC = Survey of Consumer Sentiments

tn = percentage of total nonresponse

SLC = Survey of Living Conditions

rn = percentage of refusals

NTS = National Travel Survey

HS = Holiday Survey

As mentioned before a relationship between the variable under investigation and the response behaviour reduces the value of the conclusions of the survey. The existence of such relationships is not rare, as will be illustrated in the following examples. If the aim of the survey is to measure in which way people spend their spare time, then the reason of nonresponse "not at home" is rather annoying since these people are probably spending their (spare) time somewhere else. The same applies for the survey on the number of hours people watch television: the not-at-homes (in the evening) are probably not watching television. One of the aims of the Housing



Demand Survey is to measure the frequency with which people move to other houses. As there is a considerable amount of nonresponse due to moving (the sampling unit is a person), the estimate for the total population will be biased. A number of surveys show that unmarried people have a smaller response rate. If there is a relationship between marital status and the variable under investigation then estimates will be wrong in this case too.

### 2.3 Response Models

The first requirement in the development of theories for the treatment of nonresponse is the formulation of a mathematical model, which describes the way in which nonresponse is generated. Two models appear frequently in the literature. They are denoted here by "random response model" and "fixed response model".

According to the random response model every element in the population has a certain (unknown) probability of response. These response probabilities are not necessarily the same for every element. When the interviewer contacts the person to be questioned the probability mechanism is activated and determines whether or not the person responds.

The fixed response model assumes the existence of two strata in the population: a stratum of potential respondents and a stratum of potential non-respondents. Size and content of each stratum is not known beforehand. They are determined by the specification of the survey (aim, type of questions, interviewing techniques, interviewers, period of field work, etc.). Disregarding the two strata a sample is selected from the population. Consequently the number of respondents is a random variable in both the random response model and the fixed response model.

If instead of sampling complete enumeration would take place then in the case of random response model the determination of respondents would still be a random process whereas in the case of the fixed response model this would be fixed. There is, however, a certain resemblance between the two models. Assuming the existence of two stochastic mechanisms, the

sampling mechanism and the response mechanism, both models differ only in the order in which the mechanisms are applied: In the fixed response model first the response mechanism is activated for each element in the population. This determines the two strata. Then the sample is selected. In the random response model first the sample is selected. Then the response mechanism is activated for each selected element.

The random response model offers the opportunity to estimate response probabilities. These estimated response probabilities can be used in adjustment procedures, or they can be connected to personal characteristics. The fixed response models generally results in easier formulae. The theory, developed within this model, is conditional on the realized response and non-response strata. Consequently the accuracy of the estimates can be computed, but the accuracy of the estimation method can not be determined. Due to this last argument research is focussed on the random response model.

### 3. SELECTION OF AUXILIARY VARIABLES

#### 3.1 Auxiliary Variables

It is important to discover a possibly existing relationship between the variable under investigation and the response behaviour. It is, however, not possible to determine such a relationship using the sample data, since the values of the variable under investigation are not known for the nonrespondents. To be able to say something about nonrespondents there must be information available about them. One source of information about the non-response is formed by auxiliary variables. Auxiliary variables are defined as variables which can be measured for both respondents and nonrespondents. Two types of auxiliary information can be distinguished:

- (1) Information which can be collected by the interviewer without a face-to-face interview. Among the information, obtained in this way, are type of town, type of housing, (approximate) year of construction of the housing and social status of the neighbourhood.
- (2) Information which can be obtained from administrative records. Typical examples are age, sex and marital status.

Analysis of the relationship between auxiliary variables and the response behaviour provides insight in the group of people which do not respond. It may give additional information about the relationship between the variable under investigation and the response behaviour. Auxiliary variables showing a clear relationship with the response behaviour play an important role in adjustment procedures, to be discussed later.

It is assumed that auxiliary variables are nominal variables, i.e. different values have no other meaning than to distinguish between different groups. Arithmetic operations on these values, which in fact are only labels, are not allowed. The assumption that the variables are nominal is in practice not a restriction. Many variables are nominal and other types of variables can easily be re-expressed in terms of nominal variables. As an example of the available amount of auxiliary information, the auxiliary variables of the Housing Demand Survey 1977/1978 is listed below.

- |                             |   |
|-----------------------------|---|
| (1) year of birth           | (7) number of floors in the housing     |
| (2) sex                     | (8) year of construction of the housing |
| (3) marital status          | (9) municipality                        |
| (4) size of the family      | (10) quarter of town                    |
| (5) structure of the family | (11) degree of urbanization             |
| (6) type of housing         |   |

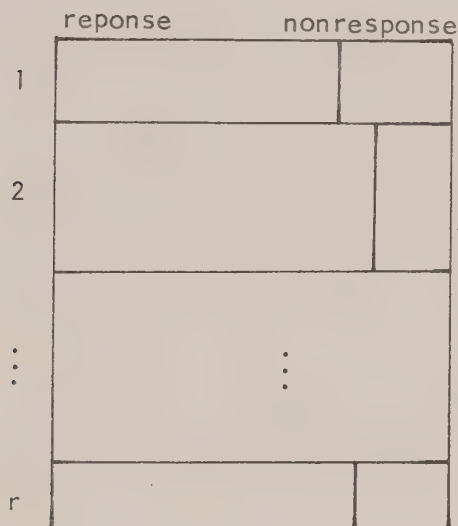
### 3.2 Graphical Methods

As a preliminary tool in the selection of auxiliary variables graphical methods have been developed. The advantage of graphical methods is clear. They bring out hidden facts and relationships and can stimulate as well as aid the analysis. They often offer a more complete and better balanced understanding than could be obtained from tabular or textual forms of presentation. Furthermore the visual relationships in the plots are more clearly grasped and more easily remembered. (See Schmid (1954).) Two simple graphical devices are presented in the next sections: the box-plot and the windmill-plot.

### 3.2.1 The box-plot

The box-plot can be seen as a generalization of a histogram or bar chart. The name of the box plot is derived from its form (see fig. 2).

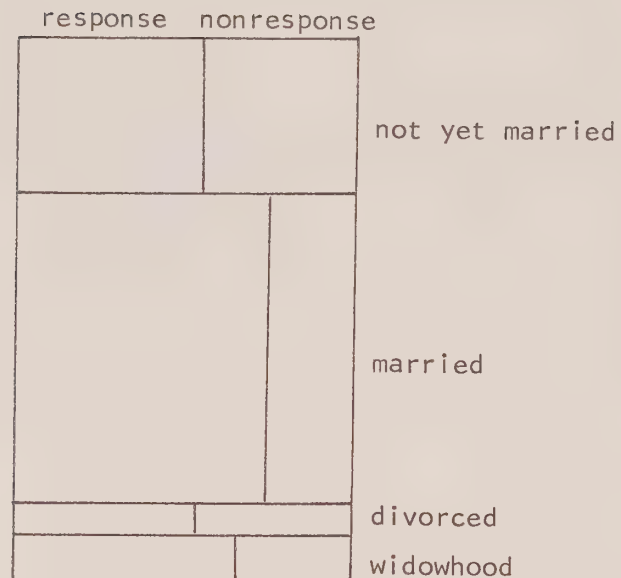
FIGURE 2. THE BOX-PLOT



A rectangle of standard width and a height proportional to the sample size represents the sample. The rectangle is divided in a number of layers (the categories of the auxiliary variable). The height of a particular layer is proportional to the number of sample elements in the corresponding category. Each layer is divided by a vertical line in a left-hand part (the response) and a right-hand part (the nonresponse). The areas of these two parts are proportional to the amounts of response and nonresponse in the particular category. Fig. 3 contains an example of a box-plot. The data originate from the Housing Demand Survey 1977/1978 as far as it concerns Amsterdam. The auxiliary variable is the marital status of the person in the sample.



FIGURE 3. BOX-PLOT OF MARITAL STATUS IN AMSTERDAM IN THE HOUSING DEMAND SURVEY 1977/1978.



A number of aspects may be worth paying attention to:

- (1) The heights of the layers indicate to what extent categories contribute to the sample. Clearly a large part of the people is married. The smallest category is the category of people who are divorced.
- (2) The extent of the nonresponse can be read from the distance of the vertical dividing lines to the right-hand side of the box. In this example there obviously is a considerable amount of nonresponse.
- (3) If all dividing lines form approximately a straight line there is no relationship between response behaviour and the auxiliary

variable. Clearly, in this situation there exists a relationship: Married people respond better than other people. Response is bad in the group of unmarried and divorced people.

More about the box plot can be found in Bethlehem & Kersten (1981).

### 3.2.2 The Windmill-Plot

The windmill-plot is a graphical representation of the results of correspondence analysis. Correspondence analysis is a technique for the analysis of associations in two-way tables. (See e.g. Benzecri (1976).). A geometrical representation of the rows (the categories of the vertically tabulated variable) and the columns (the categories of the horizontally tabulated variable) is constructed. This geometrical representation contains all the information concerning the associations in the table. By means of a scaling procedure rows and columns are assigned values in such a way that the correlation coefficient, computed by using these values, is maximized. To each cell in the table there correspond two scale values: a row-value and a column-value. When these values are conceived as coordinates, a plot of the table can be constructed. In this plot all points form an unequally spaced grid. Such a plot may not be easy to interpret. To simplify interpretation regression lines are plotted instead of the points themselves. Due to the special properties of the scale values the regression line to explain y-values from the x-values in the plot has the simple form

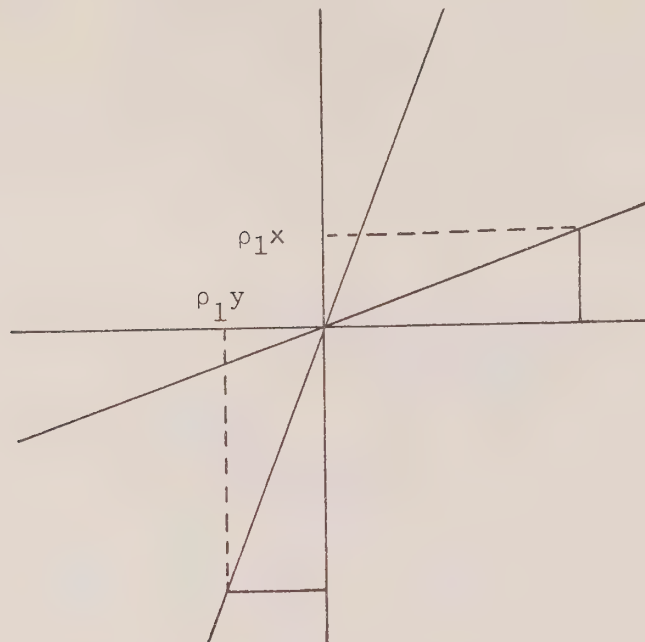
$$y = \rho_1 x \quad (1)$$

and the regression line to explain the x-values from the y-values has the form

$$x = \rho_1 t \quad (2)$$

where  $\rho_1$  is the maximized correlation coefficient. By plotting both regression lines the result is the windmill-plot, see fig. 4.

FIGURE 4. THE WINDMILL-PLOT



A number of aspects may be worth noting:

- (1) The origin represents both marginal distributions of the table
- (2) Scale values close to the origin point at categories which resemble the marginal distribution and thus have a regular behaviour. Far out scale values indicate differently behaving categories.
- (3) The relationship between the two variables is strong if the two regression lines are near the  $45^\circ$ -line.
- (4) Projection of a differently behaving category of one variable via the regression line on the axis of the other variable provides a clue about the dependencies of the categories of the variables.

The plot as described above can not account for all the information in the table. It explains as much as is possible in a two-dimensional plot. Conditionally on the first plot a second plot can be constructed, which

accounts for as much as is possible of the information not yet explained. If necessary even more plots can be constructed, but preferably one plot is sufficient to explain the major part of the associations.

A total of  $s$  of such plots can be made, in which  $s$  is one less than the minimum of the number of rows and the number of columns. Let  $\rho_1, \rho_2, \dots, \rho_s$  be the maximized correlation coefficients. Since

$$\sum_{i=1}^s \rho_i^2 = X^2/N, \quad (3)$$

where  $X^2$  is the chi-square test statistics for the table and  $N$  the general total,

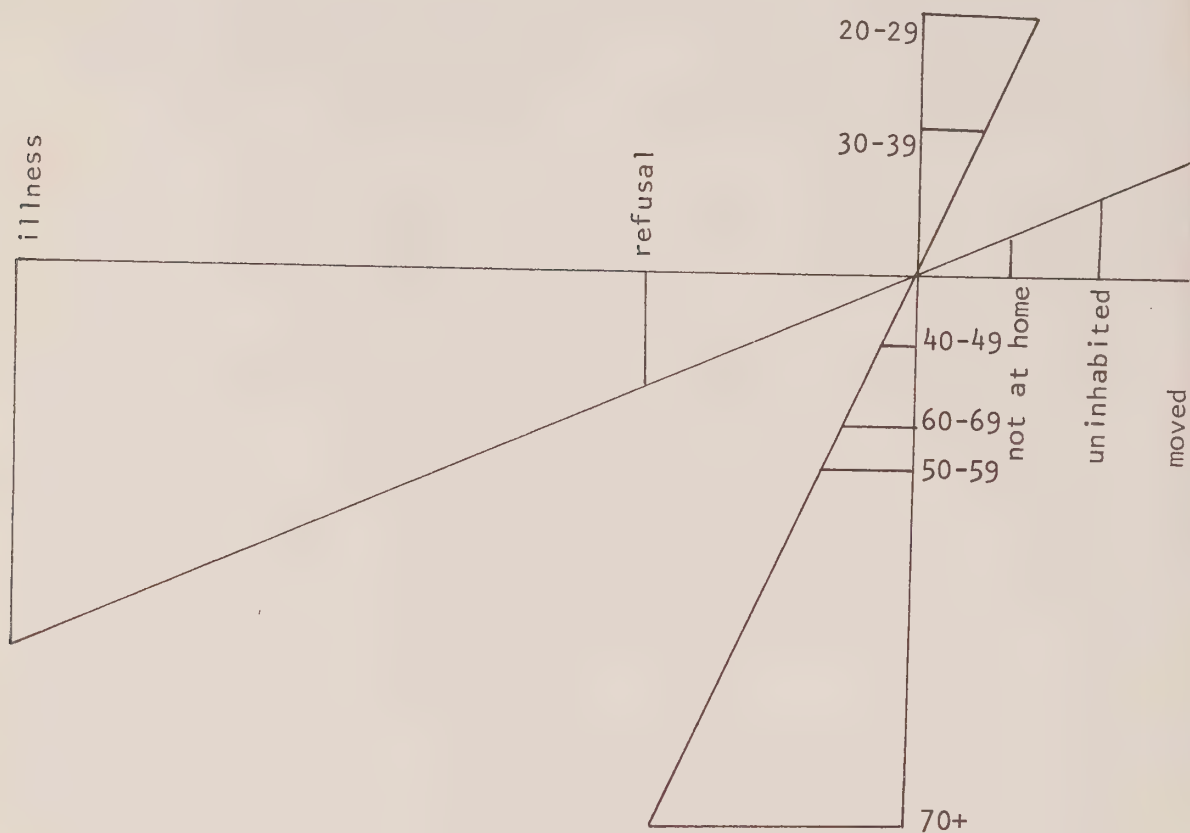
$$\tau_i = N\rho_i^2/X^2 \quad (4)$$

is a measure of the amount of information explained by the  $i$ -th plot ( $i=1, 2, \dots, s$ ).

Fig. 5 contains the first windmill-plot for the variables age (six categories) and type of nonresponse (five categories) of the Housing Demand Survey 1977/1978 as far as it concerns Amsterdam.



FIGURE 5: WINDMILL-PLOT OF AGE BY TYPE OF NONRESPONSE IN AMSTERDAM IN THE HOUSING DEMAND SURVEY 1977/78



It contains about 88% of the information about associations in the table ( $\tau_1 = 0.88$ ). The main reasons for nonresponse of the old people are refusal and illness. In case of young people the nonresponse is the result of the impossibility of making contact: uninhabited, not at home and moved. More about the application of correspondence analysis can be found in Bethlehem & Kersten (1980).

### 3.3 Other selection methods

There are many other, mainly nongraphical, method to determine the association between auxiliary variables and the response behaviour. Much about association in contingency tables can e.g. be found in Bishop, Fienberg & Holland (1975).

A popular method for the selection of the most important auxiliary variables is AID (Automatic Interaction Detection), described by Morgan & Sonquist (1963). In a stepwise process those auxiliary variables are determined which can explain as much as possible of the variance of the binary response variable. There are disadvantages which make reliable application of this method doubtful. As the selection process proceeds in a stepwise fashion there is no guarantee that the optimal solution will be found. Because there is no stopping rule based on a statistical model this sense the result is rather arbitrary. Further research in this field is necessary (see e.g. Kass (1980)).

## 4. REDUCTION OF NONRESPONSE BIAS BY SUBGROUP WEIGHTING

When a relationship is found or suspected between the variable under investigation (Y) and the response behaviour (R) measures have to be taken in order to reduce the nonresponse bias. In this section a number of adjustment procedures are discussed which are based on subgroup weighting. Attention is focussed on estimating the population mean of Y.

It can be shown that the bias, introduced by only using response values, is proportional to the covariance between Y and R. If it would be possible to divide the population in a number of subgroups in each of which the covariance is neglectable, then (nearly unbiased) estimates of the subgroup means can be combined into a (nearly unbiased) estimate of the population mean.

Let the finite population consist of N elements  $U_1, U_2, \dots, U_N$  with Y-values  $Y_1, Y_2, \dots, Y_N$ . From this population a simple random sample  $u_1, u_2, \dots, u_n$

(stochastic variables are underlined) of size  $n$  is selected without replacement. The corresponding  $y$ -values are  $y_1, y_2, \dots, y_n$  and the response behaviour is indicated by  $r_1, r_2, \dots$  ( $r_i = 1$  indicating response and  $r_i = 0$  nonresponse). In fact  $y_i$  can only be observed for those sample elements  $u_i$  for which  $r_i = 1$ . The  $m$  responding elements are denoted by  $u_{-1}, u_{-2}, \dots, u_{-m}$  ( $m = r_1 + r_2 + \dots + r_n$ ), with  $y$ -values  $y_1^*, y_2^*, \dots, y_m^*$ .

Let  $X$  be an auxiliary variable inducing a division of the population in  $H$  subgroups with sizes  $N_1, N_2, \dots, N_H$ . In subgroup weighting first of all in each subgroup  $h$  an estimator  $\bar{y}_h^*$  for the subgroup mean is computed:

$$\bar{y}_h^* = \frac{1}{m_h} \sum_{i=1}^{m_h} y_{hi}^*, \quad (h = 1, 2, \dots, H) \quad (5)$$

where  $y_{h1}^*, y_{h2}^*, \dots, y_{hm_h}^*$  are the values of the  $m_h$  responding elements in subgroup  $h$ . The subgroup estimators  $\bar{y}_1^*, \bar{y}_2^*, \dots, \bar{y}_H^*$  are combined into a population estimators  $\bar{y}^*$ .

$$\bar{y}^* = \sum_{h=1}^H w_h \bar{y}_h^* \quad (6)$$

The type of estimator is determined by the available amount of information about the weights  $w_1, w_2, \dots, w_H$ .

If the sizes  $N_1, N_2, \dots, N_H$  of the subgroups are known the situation is equivalent to poststratification. (See e.g. Holt & Smith (1979).) The weights are not random but fixed quantities:

$$w_h = \frac{N_h}{N} \quad (h = 1, 2, \dots, H) \quad (7)$$

If these sizes are not known they can be estimated by

$$w_h = \frac{n_h}{n}, \quad (h = 1, 2, \dots, H) \quad (8)$$

where  $n_h$  is the number of sample elements in subgroup  $h$  ( $n = n_1 + n_2 + \dots + n_H$ ).

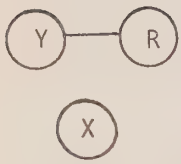
In an intermediate situation where two auxiliary variables  $X_1$  and  $X_2$  are used and only the marginal totals of the two variables are known, a raking procedure can be applied to estimate the weights (see e.g. Chapman (1976)). Suppose  $X_1$  induces  $G$  groups and  $X_2$  induces  $H$  groups. Crossing  $X_1$  and  $X_2$  results in a subdivision into  $G \times H$  groups. If only the marginal totals  $N_{g+}$  ( $g=1, 2, \dots, G$ ) of  $X_1$  and  $N_{+h}$  ( $h=1, 2, \dots, H$ ) of  $X_2$  are known then by using the sample information good estimates  $N_{gh}$  of  $N_{gh}$  can be computed. The weights are then equal to

$$w_{gh} = \frac{N_{gh}}{N} \quad (g=1, 2, \dots, G; h=1, 2, \dots, H) \quad (9)$$

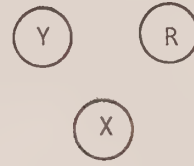
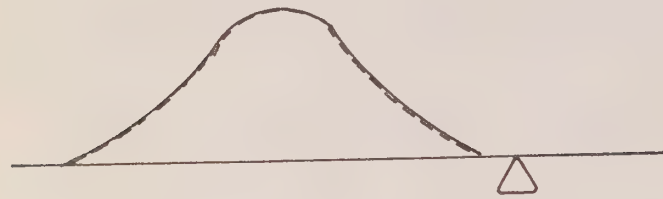
All three estimators have, when used in the same grouping situation, the same bias, but the greater the amount of available information on the subgroup sizes the smaller the variance of the estimate. Subgroup weighting has two advantages: reduction of the variance of the estimate and reduction of the response bias. The most extreme possibilities are illustrated in fig. 6. If two variables are connected it means that they have a strong correlation.



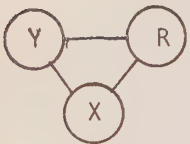
FIG. 6. VARIANCE AND BIAS OF ESTIMATORS BEFORE AND AFTER SUBGROUP WEIGHTING



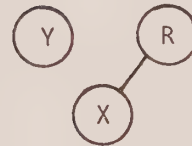
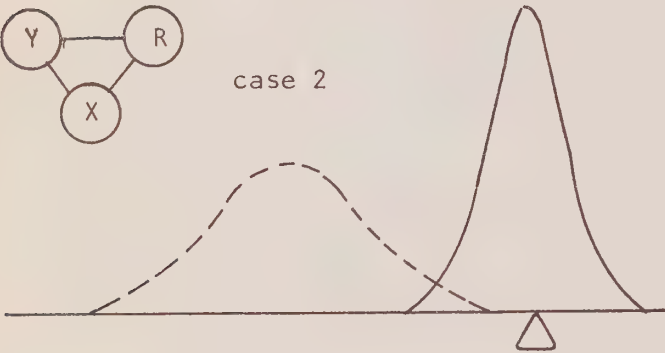
case 1



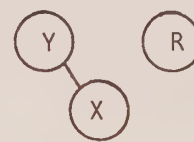
case 3



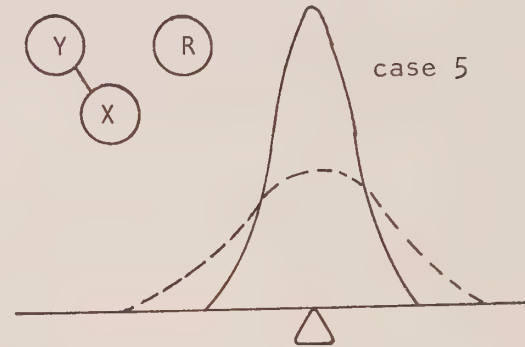
case 2




case 4



case 5



-  parameter to be estimated
- before subgroup weighting
- after subgroup weighting
- Y variable under investigation
- R response variable
- X auxiliary variable

A number of conclusions can be drawn:

- (1) If nonresponse bias exists subgroup weighting is significant when X and R are correlated (case 2). Both bias and variance are reduced.
- (2) If no nonresponse bias exists a correlation between X and R has no effect (case 4). Only correlation between X and Y reduces the variance (case 5).

Because the data on the nonrespondents are missing, it is impossible to use the remaining data to find an auxiliary variable X which is highly correlated with Y. It is, however, possible to use this data to look for auxiliary variables which are highly correlated with the response variable R. If such a variable has been found, application of it in subgroup weighting will reduce the nonresponse bias (if it exists), but not always the variance.

## 5. Other adjustment methods

Several other adjustment methods appear in the literature. Several of them will be discussed in this section. Some of them need further research to establish their merits.

### 5.1 No adjustment

In some situations no adjustment is necessary. If it appears that no relationship exists between the variable under investigation and the response behaviour the response can be considered as a random sample from the population. Also if statements are restricted to the population of potential respondents no correction is necessary. In all other situations no adjustment is only justified if the category "nonresponse" is included in all tables in publications.

### 5.2. Imputation

Imputation procedures solve the problem of missing observations due to nonresponse by substitution of values in the records of the nonrespondents. In "hot deck" imputation data are taken from respondents of the current survey, while in "cold deck" imputation data are taken from a previous survey. If the response structure of previous and current survey resemble each other the results of cold deck imputation and hot deck imputation will roughly be the same. Imputation can be carried out in several ways. Some of them are:

- (1) imputation of a random respondent
- (2) imputation of the mean respondent
- (3) imputation of a random respondent within the same subgroup
- (4) imputation of the mean respondent within the same subgroup
- (5) imputation of a value obtained by fitting a model
- (6) imputation of upper or lower bounds

Procedures (1) and (2) do not reduce the bias. Procedures (3) and (4) resemble subgroup weighting. The effect of procedure (5) depends strongly on the fit of the model and the reasonableness of the model assumptions. Procedure (6) gives insight in how bad things could be if no adjustment would take place.

### 5.3. Adjustment for not-at-homes

The well-known method of Politz & Simmons (1949) tries to adjust for not-at-home bias by estimating the probability to find a person at home. This is performed by asking respondents e.g. how often they were at home at the time of the interview during the previous days. The at-home-probability, constructed in this way, can be used as a stratification variable. It is also worth trying to find a model which explains the relationship between the variable under investigation and the at-home-probability. Extrapolation of this model to the group of not-at-homes may provide more information about this group.

#### 5.4. Adjustment for refusers

It is possible to measure the willingness of people to co-operate in the survey (see Van Tulder (1977)). Using this information a procedure analogous to adjustment for not-at-homes can be carried out. Furthermore the willingness to co-operate is a measure for the survey climate. The construction of a scale to obtain this information will probably be somewhat more difficult then in the case of not-at-home adjustment.

#### 5.5 Double sampling

In order to get more information about nonrespondents Hansen & Hurwitz (1946) propose selecting a sample from the nonrespondents. Specially trained interviewers try as yet to obtain (part of) the missing information. Time and money constraints often prevent application of double sampling.

#### 5.6. The principal question

If the method of Hansen & Hurwitz is too expensive the principal question procedure may offer a substitute. In many surveys there often is one important basic question around which the survey has been constructed. If during the field work problems are met with completing the whole questionnaire, the interviewer may try to get an answer on only the principal question. This may even be tried afterwards by letter or by telephone. This technique will shortly be tried out in one of the surveys of the CBS.

#### 6. Conclusions

In view of the rise in nonresponse rates during the past years it is important to carry out thorough research on the impact of nonresponse on the quality of the survey.

Quite a few adjustment procedures appear in literature, which all aim at reduction of the nonresponse bias. A comparative study of these procedures has to provide decisive answers about their merits.



The large differences which exist with regard to objective, design and execution of surveys prevent correct interpretation of differences in nonresponse figures. It is therefore necessary to create a theoretical framework which allows proper comparison.

Of course reduction of nonresponse during the field work will remain an important topic.

REFERENCES

- [1] Benzécri, J.P. (1976) L'Analyse des Données. Dunod, Paris.
- [2] Bethlehem J.G. and Kersten, H.M.P. (1981), "Graphical Methods in Non-Response Analysis and Sample Estimation", Staatsuitgeverij, The Hague.
- [3] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge.
- [4] Chapman, D.W. (1976), "A Survey of Non-Response Imputation Procedures", of the American Statistical Association, Social Statistics Section, 245-251.
- [5] Hansen, M.H. and Hurwitz, W.N. (1946), "The Problem of Non-Response in Sample Surveys", Journal of the American Statistician, 41, 517-529.
- [6] Holt, D. and Smith, I.M.F. (1979), "Post Stratification", Journal of the Royal Statistical Society, series A, 142, 33-46.
- [7] Kass, G.V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Applied Statistics, 29, 199-217.
- [8] Kish, L. (1967), Survey Sampling, Wiley, New York.
- [9] Morgan, J.N. and Sonquist, J.A. (1963), "Problems in the Analysis of Survey Data", Journal of the American Statistical Association, 58, 415-434.
- [10] Politz, A. and Simmons, W. (1949), "An Attempt To Get the Not-At-Homes into the Sample Without Callbacks", Journal of the American Statistical Association, 44, 9-31.
- [11] Schmid, C.F. (1954), Handbook of Graphical Presentation, Ronald Press, New York.

- [12] Tulder, J.J.M. van (1977), "Op de grens van non-response",  
Jaarboek van de Nederlandse Vereniging van Marktonderzoekers,  
1977, 43-52.
  
- [13] Widdershoven, M. and Berg, J. van den (1980), "Non-respons  
bij twee 'persoons- en gezins-enquêtes'", CBS-Select 1,  
357-365. Staatsuitgeverij, The Hague.

ON THE VARIANCES OF ASYMPTOTICALLY  
NORMAL ESTIMATORS FROM COMPLEX SURVEYSDavid A. Binder<sup>1</sup>

The problem of specifying and estimating the variance of estimated parameters based on complex sample designs from finite populations is considered. The results of this paper are particularly useful when the parameter estimators cannot be defined explicitly as a function of other statistics from the sample. It is shown how these results can be applied to linear regression, logistic regression and loglinear contingency table models.

## 1. INTRODUCTION

In recent years, there has been an increasing demand for using survey data to estimate the parameters of traditional models such as regression parameters, discriminant functions, logit and probit parameters and others. However, for many such surveys, the primary objectives of the survey is the estimation of population or sub-population means, totals, trends and so on. For this reason and because of operational considerations, the survey design is often not a simple random sample, but is more typically stratified and often multi-stage with possibly unequal probabilities at certain stages of sampling.

Because of this, there has been much discussion (see, for example, Sarndal;1978) on whether the sampling weights should be used in making inferences about these model parameters. The answer seems to depend on whether a superpopulation model is appropriate for all population units. If this is the case, the inference on the superpopulation parameters is often the primary concern. This leads to model-based inference, where, for a given sample, the inferences do not depend on the sampling weights.

---

<sup>1</sup> D.A. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada.

The question that comes to mind is: If the superpopulation model is not appropriate, what parameters are we estimating? It must be recognized that for many studies, particularly in the social sciences, the model (e.g. linear regression) is only a convenient approximation of the real world and the parameters of that model (e.g. correlations and partial correlations) are often used to understand the approximate interdependencies of the variables rather than having a particular scientific interpretation. Therefore, the parameters we are estimating do not necessarily refer to a true superpopulation model, but are of a more descriptive nature.

In this paper, we adopt the view that we are interested in making inferences about these "descriptive" parameters of the population. For example, suppose  $\underline{X}$  and  $\underline{Y}$  are  $N \times p$  and  $N \times 1$  matrices respectively, where each row of  $\underline{X}$  and  $\underline{Y}$  corresponds to a different individual of the population. We are interested in the descriptive parameter,  $\underline{B}$ , a  $p \times 1$  vector satisfying the equations:

$$\underline{X}^T \underline{X} \underline{B} = \underline{X}^T \underline{Y} \quad (1.1)$$

This view of descriptive parameters is the same as that taken by Frankel (1971) and Kish and Frankel (1974).

The usual estimation of such parameters normally takes into account the sampling weights. If we denote by  $\pi_i$  the probability that the  $i$ -th unit in the sample is sampled and let  $\underline{\Pi} = \text{diag} (\pi_1, \dots, \pi_n)$ , then the weighted parameter estimate for  $\underline{B}$  satisfies:

$$\underline{x}^T \underline{\Pi}^{-1} \underline{x} \underline{B} = \underline{x}^T \underline{\Pi}^{-1} \underline{y}, \quad (1.2)$$

where  $\underline{x}$  and  $\underline{y}$  are  $n \times p$  and  $n \times 1$  matrices respectively, the rows of which correspond to the sampled rows of  $\underline{X}$  and  $\underline{Y}$ .

Suppose, now, an estimator of a population parameter can be expressed as:

$$\hat{\theta} = g(z_1, \dots, z_k), \quad (1.3)$$



where  $E(z_i) = Z_i$ . Here,  $\hat{\theta}$  is an estimator of  $g(Z_1, \dots, Z_k)$ . Following Tepping (1968) and Woodruff (1971), a Taylor series expansion for  $\hat{\theta}$  yields:

$$V[\hat{\theta}] \doteq V\left[\sum_{i=1}^k \left(\frac{\partial g}{\partial Z_i}\right)(z_i - Z_i)\right] . \quad (1.4)$$

These formulae are exemplified for estimation of regression coefficients (1.1) by Tepping (1968). However, the expressions resulting from (1.4) for the variances of the regression coefficients are somewhat complicated compared to those derived by Fuller (1975).

In this paper we consider parameters which are not defined through an explicit equation such as (1.3), but instead are defined implicitly as  $U(\tilde{Z}, \tilde{\theta}) = 0$ . A simple example showing the distinction would be the ratio parameter:

$$R = \frac{\sum Y_k}{\sum X_k} ,$$

which could also be defined implicitly as:

$$\sum Y_k - R \sum X_k = 0..$$

When we deal with some models such as indirect loglinear models or logistic regression models, the parameters can be defined only through implicit relationships. The extension of Tepping's (1968) results for this case is fairly straightforward, but does not appear in its general form at present in the literature. There are, however, specific examples of its application; see, for example Fuller (1975) and Freeman and Koch (1976).

In Section 2 we give the general framework and the main results of the paper. A number of models are exemplified in Section 3.

## 2. GENERAL FRAMEWORK

### 2.1 Framework

The population units are labelled  $1, \dots, N$ . Associated with the  $i$ -th unit we have a  $q$ -dimensional data vector  $X_i$ . We have a parameter space  $\Theta \subseteq R^p$ . The parameter  $\theta_o = (\theta_{1o}, \dots, \theta_{po})$  is defined by the  $p$  equations:

$$U_i(X, \theta_o) = \sum_{k=1}^N u_i(X_k, \theta_o) - v_i(\theta_o) = 0, \quad (2.1)$$

for  $i=1, \dots, p$ . We assume that equations (2.1) define  $\theta_o$  uniquely in  $\Theta$ . We also assume that  $\partial u_i(X, \theta)/\partial \theta$  and  $\partial v_i(\theta)/\partial \theta$  exist in a neighbourhood of  $\theta_o$ . A simple example of (2.1) is where  $\theta_o$  is a population total, and we have  $U(X, \theta_o) = \sum_{k=1}^N X_k - \theta_o$ . Here,  $u(X_k, \theta_o) = X_k$  and  $v(\theta_o) = \theta_o$ .

We select a sample of the units, according to some probability distribution defined on the set of all non-empty subsets of  $\{1, \dots, N\}$ . We denote by  $x_1, \dots, x_n$  the selected values of  $X_1, \dots, X_N$ . We assume that for any  $\theta \in \Theta$ , we can construct a consistent, asymptotically normal estimator of  $U_i(X, \theta)$ . We denote this estimator by  $\hat{U}_i(x, \theta)$ . For example, for many without replacement sampling schemes,

$$\hat{U}_i(x, \theta) = \sum_{k=1}^n u_i(x_k, \theta)/\pi_k - v_i(\theta) \quad (2.2)$$

will be a consistent asymptotically normal estimator, where  $\pi_k$  is the probability of inclusion for the  $k$ -th unit.

We let  $\sigma_{ij}(X, \theta) = \text{Cov}[\hat{U}_i(x, \theta), \hat{U}_j(x, \theta)]$ . For example, for estimator (2.2), we have:

$$\sigma_{ij}(X, \theta) = \sum_{k=1}^N \sum_{\ell=1}^N u_i(X_k, \theta) u_j(X_\ell, \theta) (\pi_{k\ell} - \pi_k \pi_\ell) / \pi_k \pi_\ell, \quad (2.3)$$

where  $\pi_{k\ell}$  is the probability that the  $k$ -th and  $\ell$ -th units in sample.

We let  $\underline{\Sigma}(\underline{X}, \underline{\theta})$  be the  $p \times p$  matrix with entries  $\sigma_{ij}(\underline{X}, \underline{\theta})$ , and  $\hat{\underline{\Sigma}}(\underline{x}, \underline{\theta})$  be a consistent estimator for  $\underline{\Sigma}$ . Now, for any given  $\underline{\theta}$ ,

$$U_i(\underline{X}, \underline{\theta}) + v_i(\underline{\theta}) = \sum_{k=1}^N u_i(\underline{x}_k, \underline{\theta}),$$

so that estimators  $\hat{U}_i(\underline{X}, \underline{\theta})$  and  $\hat{\underline{\Sigma}}(\underline{x}, \underline{\theta})$  can be specified for any design in which we can derive consistent asymptotically normal estimators of population totals and consistent estimators for the variances of the estimators of the totals.

The Horvitz-Thompson estimator for (2.3) is:

$$\sum_{k=1}^n \sum_{\ell=1}^n u_i(\underline{x}_k, \underline{\theta}) u_j(\underline{x}_\ell, \underline{\theta}) (\pi_{k\ell} - \pi_k \pi_\ell) / \pi_k \pi_\ell \pi_{k\ell}. \quad (2.4)$$

In the case of fixed sample size, the Yates-Grundy estimator of (2.3) is:

$$\sum_{k < \ell} \left[ \frac{u_i(\underline{x}_k, \underline{\theta})}{\pi_k} - \frac{u_i(\underline{x}_\ell, \underline{\theta})}{\pi_\ell} \right] \left[ \frac{u_j(\underline{x}_k, \underline{\theta})}{\pi_k} - \frac{u_j(\underline{x}_\ell, \underline{\theta})}{\pi_\ell} \right] (\pi_k \pi_\ell - \pi_{k\ell}). \quad (2.5)$$

Letting  $\underline{U}(\underline{X}, \underline{\theta})$  and  $\hat{\underline{U}}(\underline{x}, \underline{\theta})$  be the  $p$ -dimensional vectors with components  $U_i(\underline{X}, \underline{\theta})$  and  $\hat{U}_i(\underline{x}, \underline{\theta})$  respectively, we define

$$\underline{J}(\underline{X}, \underline{\theta}) = \partial \underline{U}(\underline{X}, \underline{\theta}) / \partial \underline{\theta} \quad (2.6)$$

$$\hat{\underline{J}}(\underline{x}, \underline{\theta}) = \partial \hat{\underline{U}}(\underline{x}, \underline{\theta}) / \partial \underline{\theta}, \quad (2.7)$$

where  $\underline{J}$  and  $\hat{\underline{J}}$  are  $p \times p$  partial derivative matrices. Assume that the matrices are continuous functions of  $\underline{\theta}$  and that the partial derivatives with respect to  $\underline{\theta}$  exist in a neighbourhood of  $\underline{\theta}_0$ . Also assume  $\hat{\underline{J}}(\underline{x}, \underline{\theta})$  is a consistent estimator of  $\underline{J}(\underline{X}, \underline{\theta})$ .

Our estimator for  $\underline{\theta}$  is given by  $\hat{\underline{\theta}}$ , the solution to:

$$\hat{U}_i(\underline{x}, \hat{\underline{\theta}}) = 0, \text{ for } i=1, \dots, p. \quad (2.8)$$

We assume the sample size is sufficiently large so that the solution to (2.8) is unique in  $\theta$ . We show in the next section that the covariance matrix of  $\hat{\theta}$  can be consistently estimated by:

$$[\hat{J}^{-1}(\underline{x}, \hat{\theta})] \hat{\Sigma}(\underline{x}, \hat{\theta}) [\hat{J}^{-1}(\underline{x}, \hat{\theta})]^T.$$

## 2.2 Asymptotic Theory

Following the asymptotic arguments of Madow (1948), and Hájek (1960), we consider a sequence of populations indexed by  $t$ , with sizes  $N^{(t)}$  and data  $\underline{x}^{(t)}$ . We assume  $N^{(t)} \rightarrow \infty$  as  $t \rightarrow \infty$ . For population  $t$ , we select a sample of size  $n^{(t)}$  and observe data  $\underline{x}^{(t)}$ . We let  $v^{(t)} = E(n^{(t)})$  and assume

$$\lim_{t \rightarrow \infty} v^{(t)} = \infty$$

$$\lim_{t \rightarrow \infty} (N^{(t)} - v^{(t)}) = \infty$$

For any  $\theta$  in a neighbourhood of  $\theta_0^{(t)}$  we assume

$$[v^{(t)}]^{1/2} [\hat{U}(\underline{x}^{(t)}, \theta) - U(\underline{x}^{(t)}, \theta)]/N^{(t)}$$

is asymptotically  $N[0, S(\theta)]$ , where

$$S(\theta) = \lim_{t \rightarrow \infty} [v^{(t)} \Sigma(\underline{x}^{(t)}, \theta) / \{N^{(t)}\}^2]$$

exists. We assume

$$K(\theta) = \lim_{t \rightarrow \infty} \hat{J}(\underline{x}^{(t)}, \theta)/N^{(t)} \text{ exists and also}$$

$$\text{plim } \hat{J}(\underline{x}^{(t)}, \theta)/N^{(t)} = K(\theta).$$

Also, we assume

$$\lim[\text{rank } \{\hat{J}(\underline{x}^{(t)}, \theta)\}] = \text{plim}[\text{rank } \{\hat{J}(\underline{x}^{(t)}, \theta)\}] = p.$$

We define  $\hat{\theta}^{(t)}$  to satisfy

$$\hat{U}(\underline{x}^{(t)}, \hat{\theta}^{(t)}) = 0.$$

By a Taylor series expansion, we obtain

$$\hat{U}(\underline{x}^{(t)}, \hat{\theta}^{(t)}) \doteq - \hat{J}(\underline{x}^{(t)}, \hat{\theta}^{(t)}) (\hat{\theta}^{(t)} - \theta_0^{(t)}). \quad (2.9)$$

Since the left hand side of (2.9) is asymptotically normal, we have that

$$(n^{(t)})^{1/2} (\hat{\theta}^{(t)} - \theta_0^{(t)})$$

is asymptotically  $N[0, G(\theta_0)]$ , where  $S(\theta_0) = K(\theta_0) G(\theta_0) [K(\theta_0)]^T$ .

Therefore,

$$\underline{G}(\underline{\theta}_0) = [\underline{K}^{-1}(\underline{\theta}_0)] \underline{S}(\underline{\theta}_0) [\underline{K}^{-1}(\underline{\theta}_0)]^T \quad (2.10)$$

and a consistent estimator for  $\underline{G}(\underline{\theta}_0)$  is :

$$n^{(t)} [\hat{\underline{J}}^{-1}(\underline{x}, \hat{\underline{\theta}})] \hat{\underline{S}}(\underline{x}, \hat{\underline{\theta}}) [\hat{\underline{J}}^{-1}(\underline{x}, \hat{\underline{\theta}})]^T. \quad (2.11)$$

Hence, when the functional form of  $\underline{\hat{U}}(\underline{x}, \underline{\theta})$  and  $\underline{\hat{S}}(\underline{x}, \underline{\theta})$  is specified, we need only derive the matrix  $\underline{\hat{J}}(\underline{x}, \underline{\theta}_0)$  and its estimator  $\underline{\hat{J}}(\underline{x}, \hat{\underline{\theta}})$  to use these results.

### 3. EXAMPLES

#### 3.1 Introduction

In this section we consider in detail the implication of the general formulation given in Section 2 with respect to estimating the variances of certain population parameter estimators. In particular, we discuss ratios, regression coefficients and log linear models for categorical data. Other models, such as probit models could be analyzed analogously.

In general, we use the following notation. If  $\underline{w}_1, \dots, \underline{w}_N$  are population values, with  $\underline{W} = \Sigma \underline{w}_k$ , then on selecting a sample  $\underline{w}_1, \dots, \underline{w}_n$ , we have an unbiased estimator of  $\underline{W}$  given by  $\hat{\underline{W}}$ . We let  $\underline{V}(\hat{\underline{W}})$  represent the covariance matrix for  $\hat{\underline{W}}$  and  $\underline{\hat{V}}(\hat{\underline{W}})$  a consistent estimator of  $\underline{V}(\hat{\underline{W}})$ . The particular form of this estimator will depend on the sample design; for example, multi-stage stratified, pps with replacement, etc. .

#### 3.2 Ratios

Suppose we are interested in  $R = \Sigma X_{k2} / \Sigma X_{k1}$ . We define

$$U(\underline{x}, R) = \Sigma X_{k2} - R \Sigma X_{k1}.$$

Therefore, for without replacement sampling, we have :

$$\hat{U}(\underline{x}, R) = \hat{X}_2 - R \hat{X}_1.$$



Setting  $\hat{U}(\underline{x}, \hat{R}) = 0$ , we obtain

$$\hat{R} = \hat{X}_2 / \hat{X}_1. \quad (3.1)$$

We define  $W_k = X_{k2} - R X_{k1}$ .

Since,  $J(\underline{x}, R) = -\sum X_{k1}$ , we have that  $V(\hat{R})$  is approximately  $V(\hat{W}) / (\sum X_{k1})^2$ . This is estimated by  $\hat{V}(\hat{W}) / \hat{X}_1^2$ . In the case of stratified sampling, this yields the same result as in Woodruff (1971).

### 3.3 Regression Coefficients and R

Suppose our data matrix  $\underline{X}$  is partitioned into  $[\underline{Z} | \underline{Y}]$ , the first column of  $\underline{Z}$  being the vector of 1's. The vector  $\underline{Y}$  is  $N \times 1$ . We have parameters of interest  $\theta$ ,  $\underline{B}$ , and  $R^2$  defined by:

$$U_1 = \theta - \underline{Y}^T \underline{1} = 0, \quad (3.2a)$$

$$U_2 = \underline{Z}^T \underline{Z} \underline{B} - \underline{Z}^T \underline{Y} = 0, \quad (3.2b)$$

$$U_3 = (\underline{Y}^T \underline{Y} - N^{-1} \theta^2) (R^2 - 1) + \underline{Y}^T \underline{Y} - \underline{Y}^T \underline{Z} \underline{B} = 0. \quad (3.2c)$$

Here,  $\underline{B}$  denotes the vector of regression coefficients,  $R^2$  is the coefficient of multiple determination and  $\theta$  is the total of the  $Y$ 's. We first consider the case where  $N$  is known. We let  $SSY = \underline{Y}^T \underline{Y} - N^{-1} \theta^2$ . We also define  $S_{ZZ}$  as the estimator for  $\underline{Z}^T \underline{Z}$ ,  $S_{YY}$  the estimator for  $\underline{Y}^T \underline{Y}$  and  $S_{ZY}$  the estimator for  $\underline{Z}^T \underline{Y}$ . We therefore have:

$$\hat{\theta} = \hat{Y}, \quad (3.3a)$$

$$\hat{\underline{B}} = S_{ZZ}^{-1} S_{ZY}, \quad (3.3b)$$

$$\hat{R}^2 = 1 - \frac{S_{YY} - \hat{\underline{B}}^T S_{ZY}}{S_{YY} - N^{-1} \hat{Y}^2}, \quad (3.3c)$$

and

$$\underline{J} = \partial U(\underline{Z}, \underline{Y}, \underline{B}, R^2, \theta) / \partial (\underline{B}, \hat{R}, \theta) = \begin{bmatrix} \underline{0}^T & 0 & 1 \\ \underline{Z}^T \underline{Z} & 0 & 0 \\ -\underline{Y}^T \underline{Z} & SSY & 2\bar{Y}(1-R^2) \end{bmatrix},$$

where  $\bar{Y} = \theta/N$ .

Therefore,

$$\tilde{J}^{-1} = \begin{bmatrix} 0 & (\tilde{Z}^T \tilde{Z})^{-1} & 0 \\ -2\tilde{Y}(1-R^2)/SSY & \tilde{B}^T/SSY & 1/SSY \\ 1 & 0^T & 0 \end{bmatrix}.$$

Now, letting  $\tilde{W}_k^T(B) = (Z_{k1} e_k, \dots, Z_{kp} e_k)$ , where  $e_k = Y_k - \sum_j Z_{kj} B_j$ , we obtain:

$$\tilde{V}[\hat{B}] \doteq (\tilde{Z}^T \tilde{Z})^{-1} \tilde{V}[\hat{W}(B)] (\tilde{Z}^T \tilde{Z})^{-1}. \quad (3.4)$$

This is a direct consequence of (2.10). Note that the set of  $\tilde{W}_k(B)$  vectors corresponds to  $U_2$  in (3.2b). Fuller (1975) obtains the same result for stratified or two-stage stratified sampling.

To estimate (3.4) we use:

$$\hat{\tilde{V}}[\hat{B}] = \hat{S}_{ZZ}^{-1} \hat{\tilde{V}}[\hat{W}(\hat{B})] \hat{S}_{ZZ}^{-1}.$$

We can also estimate the variance of  $\hat{R}^2$ . If  $\tilde{W}_k^T(B, R^2) = [Y_k, Z_{k1} e_k, \dots, Z_{kp} e_k, Y_k (\sum_j Z_{kj} B_j - R^2 Y_k)]$  and  $\tilde{c}^T = [-2\hat{Y}(1-\hat{R}^2)/N, \hat{B}^T, 1]/(S_{YY} - N^{-1} \hat{Y}^2)$ , we obtain:

$$\hat{\tilde{V}}[\hat{R}^2] \doteq \tilde{c}^T \hat{\tilde{V}}[\hat{W}(\hat{B}, \hat{R}^2)] \tilde{c}. \quad (3.5)$$

For the case where  $N$  is unknown (e.g. the primary sampling units are geographic areas), we have the additional equation:

$$U_4 = N - \sum 1. \quad (3.6)$$

Adding the appropriate row and column to  $\tilde{J}$  and inverting, we obtain the following results for estimating  $V[\hat{R}^2]$ .

We let

$$\tilde{W}_k^T(B, R^2) = [Y_k, Z_{k1} e_k, \dots, Z_{kp} e_k, Y_k (\sum_j Z_{kj} B_j - R^2 Y_k), 1]$$

and

$$\tilde{c}^T = [-2\hat{Y}(1-\hat{R}^2)/\hat{N}, \hat{B}^T, 1, \hat{Y}^2(1-\hat{R}^2)/\hat{N}^2]/(S_{YY} - \hat{N}^{-1} \hat{Y}^2).$$

We then have  $\hat{V}[\hat{R}^2]$  is given by (3.5) for these new values of  $\tilde{w}_k(B, R^2)$  and  $\tilde{c}$ .

### 3.4 Logistic Regression

As in the previous section, we assume the data matrix  $X$  can be partitioned into  $[Z|Y]$ , but now  $Y$  is a vector of 0's and 1's. In the traditional statistical framework, the logistic regression model for  $Y$  conditional on  $Z$  asserts that  $Y_1, \dots, Y_N$  are independent with  $\Pr(Y_k=1) = p_k(\beta)$ , where :

$$p_k(\beta) = \frac{\exp(\beta^T \tilde{z}_k)}{1 + \exp(\beta^T \tilde{z}_k)} \quad (3.7)$$

Letting  $\tilde{B}$  be the maximum likelihood estimator for  $\beta$ , we have that  $\tilde{B}$  satisfies

$$\tilde{U} = \tilde{Z}^T P(\tilde{B}) - \tilde{Z}^T Y = 0, \quad (3.8)$$

where  $P(\tilde{B})^T = [p_1(\tilde{B}), \dots, p_N(\tilde{B})]$ .

For a given finite population, we define  $\tilde{B}$  as our parameter of interest.

We let  $\tilde{C}(\tilde{B})$  be our estimate for  $\tilde{Z}^T P(\tilde{B})$  and  $\tilde{S}_{ZY}$  our estimate for  $\tilde{Z}^T Y$ . Therefore,  $\hat{\tilde{B}}$  satisfies  $\tilde{C}(\hat{\tilde{B}}) = \tilde{S}_{ZY}$ . These equations must be solved iteratively in general. We also have

$$\tilde{J} = \frac{\partial \tilde{U}}{\partial \tilde{B}}.$$

The  $(i,j)$ th component of  $\tilde{J}$  is  $\sum_k Z_{ki} Z_{kj} p_k(\tilde{B}) [1-p_k(\tilde{B})]$ . We denote the estimator of  $\tilde{J}$  by  $\hat{\tilde{J}}$ .

To estimate the variance of  $\hat{\tilde{B}}$ , we let

$$\tilde{w}_k^T = (Z_{k1} \hat{e}_k, \dots, Z_{kr} \hat{e}_k)$$

where  $\hat{e}_k = p_k(\hat{\tilde{B}}) - Y_k$ . The estimator for  $\tilde{V}[\hat{\tilde{B}}]$  is given by :

$$\hat{\tilde{J}}^{-1} \hat{\tilde{V}}(\hat{\tilde{w}}) \hat{\tilde{J}}^{-1}.$$

### 3.5 Loglinear Models for Categorical Data

Suppose that each member of the population belongs to exactly one of  $q$  distinct categories. Associated with category  $i$  we have an  $r \times 1$  vector  $\underline{a}_i$  such that the proportion of individuals in the  $i$ -th category is approximately

$$p_i(\underline{\beta}) = \frac{\exp(\underline{a}_i^T \underline{\beta})}{\sum_j \exp(\underline{a}_j^T \underline{\beta})}.$$

We let  $\underline{p}(\underline{\beta})^T = [p_1(\underline{\beta}), \dots, p_q(\underline{\beta})]$  and  $\underline{N}^T = (N_1, \dots, N_q)$ , where  $N_i$  is the number of individuals in the  $i$ -th category. Now, if the population were generated from a multinomial distribution with probabilities  $\underline{p}(\underline{\beta})$ , the maximum likelihood estimator for  $\underline{\beta}$ , given by  $\underline{B}$ , satisfies:

$$\underline{U} = \underline{A}^T \underline{N} - [\underline{A}^T \underline{p}(\underline{B})] \underline{1}^T \underline{N} = 0,$$

where  $\underline{A}$  is a  $q \times r$  matrix with  $i$ -th row being  $\underline{a}_i^T$ . We consider  $\underline{B}$  as our parameter of interest for any given finite population.

We let  $\hat{\underline{N}}$  be a consistent asymptotically normal estimator of  $\underline{N}$ , with variance-covariance matrix  $\underline{V}[\hat{\underline{N}}]$  and estimated matrix  $\hat{\underline{V}}[\hat{\underline{N}}]$ . Our estimator,  $\hat{\underline{B}}$ , satisfies:

$$\underline{A}^T \hat{\underline{N}} - [\underline{A}^T \underline{p}(\hat{\underline{B}})] \underline{1}^T \hat{\underline{N}} = 0. \quad (3.9)$$

This estimator was suggested by Freeman and Koch (1976). It may be less efficient than Imrey, Koch and Stokes (1981, 1982) functional asymptotic regression methodology; however, we need not calculate all the components of  $\hat{\underline{V}}[\hat{\underline{N}}]$  to apply (3.9).

Let  $\underline{D}(\underline{B})$  be  $\text{diag}[\underline{p}(\underline{B})]$  and  $\underline{H}(\underline{B}) = \underline{D}(\underline{B}) - \underline{p}(\underline{B}) \underline{p}(\underline{B})^T$ . We have:

$$\underline{J} = \frac{\partial \underline{U}}{\partial \underline{B}} = - (\underline{1}^T \underline{N}) \underline{A}^T \underline{H}(\underline{B}) \underline{A}.$$

Therefore the asymptotic variance matrix for  $\hat{\underline{B}}$  is given by:

$$\begin{aligned} \underline{V}[\hat{\underline{B}}] &= (\underline{N}^T \underline{1})^{-2} (\underline{A}^T \underline{H}(\underline{B}) \underline{A})^{-1} \\ &\quad \underline{A}^T (\underline{I} - \underline{p}(\underline{B}) \underline{1}^T) \underline{V}[\hat{\underline{N}}] (\underline{I} - \underline{1} \underline{p}(\underline{B})^T) \underline{A} (\underline{A}^T \underline{H}(\underline{B}) \underline{A})^{-1}. \end{aligned} \quad (3.10)$$

This expression can sometimes be simplified as follows. If it can be assumed that  $\underline{N}/\underline{N}^T \underline{1} \doteq \underline{p}(\underline{B})$ , then for  $\hat{\underline{\pi}} = \hat{\underline{N}}/\hat{\underline{N}}^T \underline{1}$  we have:

$$\underline{V}[\hat{\underline{\pi}}] \doteq (\underline{N}^T \underline{1})^{-2} (\underline{I} - \underline{p}(\underline{B}) \underline{1}^T) \underline{V}[\hat{\underline{N}}] (\underline{I} - \underline{1} \underline{p}(\underline{B})^T),$$

so that

$$\underline{V}[\underline{B}] \doteq (\underline{A}^T \underline{H}(\underline{B}) \underline{A})^{-1} \underline{A}^T \underline{V}[\hat{\underline{\pi}}] \underline{A} (\underline{A}^T \underline{H}(\underline{B}) \underline{A})^{-1}. \quad (3.11)$$

We also have that the covariance matrix for  $\underline{p}(\hat{\underline{B}})$ , the estimated cell probabilities, is given by:

$$\underline{V}[\underline{p}(\hat{\underline{B}})] = \underline{H}(\underline{B}) \underline{A} \underline{V}[\hat{\underline{B}}] \underline{A}^T \underline{H}(\underline{B}).$$

The estimators of  $\underline{V}[\hat{\underline{B}}]$  and  $\underline{V}[\underline{p}(\underline{B})]$  are similar expressions, where  $\underline{N}$  and  $\underline{B}$  are replaced by  $\hat{\underline{N}}$  and  $\hat{\underline{B}}$  respectively. These assume that  $\hat{\underline{V}}[\hat{\underline{N}}]$  is readily available. For some problems where  $q$  is relatively large compared to  $r$ , it would be more efficient to proceed as follows. Let

$$\begin{aligned} Y_{ki} &= 1 \quad \text{if } k\text{-th unit in } i\text{-th category} \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

for  $k=1, \dots, N$ ;  $i=1, \dots, q$ . Let  $\underline{Y}_k^T = (Y_{k1}, \dots, Y_{kq})$ , and

$$\underline{W}_k = \underline{A}^T [\underline{I} - \underline{p}(\hat{\underline{B}}) \underline{1}^T] \underline{Y}_k.$$

We then obtain:

$$\hat{\underline{V}}[\hat{\underline{B}}] = (\hat{\underline{N}}^T \underline{1})^2 (\underline{A}^T \underline{H}(\hat{\underline{B}}) \underline{A})^{-1} \hat{\underline{V}}(\hat{\underline{W}}) (\underline{A}^T \underline{H}(\underline{B}) \underline{A})^{-1}.$$

We remark that the methodology described in this section can be readily extended to product-multinomial type models, where we have a log-linear model for  $\{N_{ij}\}$ , but the margins  $\{\sum_j N_{ij}\}$  are known.



#### 4. DISCUSSION

The techniques described in the paper have been described for some specific models; see, for example, Fuller (1975) and Freeman and Koch (1976). However, the general results are not explicitly described. Many standard statistical packages may be used for the estimation of the parameters of the models described, but the variances and tests of hypotheses given in these packages will not be valid.

The results of this paper depend on the assumption of asymptotic normality of the estimators. Empirical studies on the validity of these approximations are important.

An alternative methodology to estimating many of the parameters described here is given by Imrey, Koch and Stokes (1981, 1982). Their functional asymptotic regression methodology also falls within the general framework described here, with respect to variance derivation and estimation.

REFERENCES

- [1] Frankel, M.R. (1971), Inference from Survey Samples. University of Michigan, Ann Arbor.
- [2] Freeman, D.H. Jr., and Koch, G.G. (1976), "An Asymptotic Covariance Structure for Testing Hypotheses on Raked Contingency Tables from Complex Sample Surveys", Proc. Amer. Statist. Ass. (Social Statistics Section), Part 1, 330-335.
- [3] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankya, Series C, 37, 117-132.
- [4] Hajek, J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population", Publ. Math. Inst. Hung. Acad. Sci., 5, 361-374.
- [5] Imrey, P.B., Koch, G.G., Stokes, M.E. (1981, 1982), "Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Overview. Part II: Data analysis", International Statistical Review. To appear.
- [6] Kish, L., and Frankel, M.R. (1974), "Inference from Complex Samples", J. Roy. Statistic. Soc. B, 36, 1-22.
- [7] Madow, W.G. (1948), "On the Limiting Distribution of Estimates Based on Samples from Finite Universes", Ann. Math. Statist., 19, 535-545.
- [8] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.
- [9] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Amer. Statist. Assoc. (Social Statistics Section), 11-18.
- [10] Woodruff, R.S. (1971), "A Simple Method for Approximating the Variance of a Complicated Estimate", J. Amer. Statist. Assoc. 66, 411-414.

AN OVERVIEW OF CANADIAN HEALTH STATISTICS:  
PAST, PRESENT AND FUTURE<sup>1</sup>

Lorne Rowebottom<sup>2</sup>

The author briefly reviews the factors determining the production of health statistics in Canada, with particular attention to the different sources of data and to the long-standing co-operation among the many agencies involved in the gathering of health-related information.

Mr. Chairman, I want to express my real pleasure at being a member of this panel because of the opportunity that it affords me to congratulate Dorothy Rice and her colleagues in the National Center for Health Statistics on the occasion of the completion of 25 years of Health Surveys. We in Statistics Canada have long been admirers of NCHS and my congratulations to Dorothy are on behalf of my colleagues in Statistics Canada, particularly those in our Health Division.

Consistent, I hope, with the charge of our Chairman, I have chosen to paint with a very broad brush what seem to me to be trends and determinants of our health which might find echos in other countries and therefore be of interest to this audience.

Two data streams comprise the historic and current sources of Canada Health Statistics. The first is health institutions - predominantly hospitals, both general and mental. From them we derive statistics about a wide range of their characteristics, as well as statistics about their patents and their illnesses. Canadian hospital statistics are amongst the most detailed and comprehensive in the world.

---

<sup>1</sup> As presented at the American Statistical Association Annual Meeting in Detroit, August 1981

<sup>2</sup> Lorne Rowebottom, Assistant Chief Statistician, Institutions and Agriculture Statistics Branch, Statistics Canada.

The second stream comprises the records generated by registration of births, marriages and deaths from which we derive the critical statistics on causes of death.

A wide variety of statistics is produced from such rich data bases and some important statistics are derived from other sources, for example, those on cancer incidence, from cancer registers, and notifiable diseases. For those who are interested I have a few copies of a Directory of Health Division Information and also I would be glad to send a copy to anyone who wrote to me at Statistics Canada.

The important themes relating to these statistics that I want to touch on this morning are the following:

- First, they measure illness only when individuals seek health care from institutions.
- Secondly, they illustrate the strengths and weaknesses of statistics derived from surveys and from administrative records.
- Thirdly, they represent the availability of information which could only result from a very high degree of co-operation, sustained over a long period of time, between the central agency, federal and provincial departments of health, the institution and hospital associations, and vital statistics registers.

I will return to these three characteristics of the health statistics system: what is measured and what is not, the implications of data sources and the degree of co-operation between the players in the system.

Why have we produced what we have, rather than different products by different means? Looking back over sixty years of health statistics, I found this an interesting question. Assessing how priorities were determined is a judgemental process - just as is deciding on today's priorities. So it is my judgement that in part we responded to changing needs for statistics articulated by users and Royal Commissions, and in part we anticipated changing user needs ourselves and used existing data

sources which related to such needs, and because they represented opportunities. They were there to be utilized, like the vein of quartz that a prospector seeks and finds, or stumbles across. In part, we were driven by, and we exploited, the rapidly changing technology. In part the environment of co-operation in which we worked determined what we did. And finally in many parts the resources available to us in terms of dollars, human skills, and data handling capabilities, permitted some things and not others.

These few critical factors:

- articulated and perceived needs,
- data sources available,
- changing technology to process and to analyse data,
- co-operation between players in the system,
- budgets available,

have been the determinants of what we have done. But it will be apparent to you that they are also the determinants of what we are and will be doing.

These forces shift and come together in a changing kaleidoscope so that during one span of time one combination is dominant, to be replaced by another combination.

In Canada all have operated in such ways to bring about significant changes in our health statistics and it seems apparent that there will result even more rapid change. Changing needs should, of course, drive the system and they are in fact doing so, albeit in some respect in an erratic manner. You will recall my stating that the Canadian measurements of morbidity are largely limited to hospitalized illnesses. This has been widely recognized as a quite unacceptable state of affairs and a few years ago this dissatisfaction led to a federal decision to institute a continuing health status survey of the Canadian population. A survey was carefully planned and tested from both conceptual and methodological points of view. However, only 10 months' data were collected before government-wide budget reductions forced cancellation of the survey. The first results from the



data collected have just been published and the data base has shown signs of being a rich research source with significant decision-making implications. Of course, it suffers from the severe limitations of relating to only one point in time. It is too early to state how long it may be before a decision to reinstitute some form of the Canada Health Survey is made. However, I am optimistic that the capacity of such measurements of health status - to throw light on the effects of our lifestyles on our good health and illness, and lead to individual and collective decisions which will affect them - will not be ignored for long.

Let me turn from the area of health-related household surveys where the Canadian track record of responding to changing needs is poor, to one where we have both anticipated and responded effectively to new demands. I refer to epidemiological studies designed to enlighten the kinds of health risks resulting from exposure to various demographic, social, occupational and environmental influences. Thanks to the foresight and persistence of members of our Vital Statistics Staff working with a few other key persons both within and outside Statistics Canada, we have a computer-searchable Mortality Data Base file which includes all deaths in Canada, coded by cause of death, extending back over three decades. We also have a generalized record linkage facility which is being used to link specific exposed population groups to the mortality file. Linkages are also possible to an as yet incomplete but significant ten-year cancer incidence file.

A paper which includes a largely Canadian bibliography on this area will be given by Martha Smith, Head of Occupational and Environmental Health Research Unit, in Scotland before the end of this month. It will be available on request. (Both Martha and John Silins, Chief of our Vital Statistics and Disease Registries Section are in the audience.)

As to other data available to shape the future of Canadian Health Statistics I will only take time to mention the existence of data bases which are very large, potentially very rich, and largely unused for national statistical purposes.

They comprise the administrative records of our national medicare system which record annually in excess of 30 million incidents of primary medical care extended by physicians. We have demonstrated some of the statistical potential of these files and we are now shaping new proposals to develop their use during the next several years. Budgets are expected to be the limiting factor.

New needs should drive the system - new technology does. The influence of computers on health statistics is all-pervasive and is operating to change the availability and uses of health statistics in profound ways.

I want to comment on the use of data - in the form of statistical information, which computers have made possible - by managers, medical personnel and administrators in hospitals, local hospital districts, states, provinces, universities and associations. At federal levels, computers have changed the ways in which data are processed and statistics are used. But in many locations throughout the health community, computers have meant that data are now used for purposes of understanding, for research and for decisions, whereas in the precomputer era they were used little or not at all.

Allowing for some exaggeration - but probably not very much - it was not that long ago when national statistical agencies had almost a monopoly on large-scale data handling capability. What a contrast between then and now when large, fast, sophisticated and easily used information processing capacity is economically available to both large and small organizations. The implications are far-reaching and I suspect not yet fully perceived, but they include at least:

- The existence of many rather than few producers of statistics (many of these will perceive themselves as operators of MIS but statistics is - and will be - the game if not the name.)
- These same organizations will also be much more intensive users of statistics - particularly statistics about their own organizations or jurisdictions.

- As a result there will be greater knowledge of one's own environment.
- There will be greater independence on the part of such organizations and their need - maybe much less perceived need - to rely on others for statistics.
- This ability to utilize the information contained in the administrative records of one's own organization or jurisdiction will almost certainly reduce the tolerance for completing statistical questionnaires, with a resulting increase in the necessity to rely on administrative records. This could result in less information being available about the total environment because of the problems of data comparability between organizations and jurisdictions.

I find it difficult to forecast the impact that these changes will have on co-operation between the many players essential to development and maintenance of a comprehensive and inevitably complex system of health statistics. All I can say is that in Canada - notwithstanding substantial pressures which test and strain the system - co-operation has not diminished. In fact, the reverse is the case and on this score also I am an optimist. I think that one determinant of such co-operation is for national statistical agencies to recognize that their role must change in response to the kind of changes I have described. It is apparent to me that priorities must shift from statistical production to statistical co-ordination.

One final word about what I consider to be an overriding priority, namely, doing statistical analysis of our data bases to determine the messages that are in them, to determine their meaning and significance, and to relate them to the issues and problems confronting us.

For too long, we, at least we in Statistics Canada, have published numbers - myriads of numbers - and failed to translate them into significant indicators. We have left it to others to find the gold in the ore we have mined. I think that we and the health community have paid a high price for our failures (there have been successes) to find the gold, and even shape it into jewellery with which users would enlighten our world, not unlike the way necklaces lend radiance to those who wear them.

MODELS FOR ESTIMATION OF SAMPLING ERRORS<sup>1</sup>P.D. Ghangurde<sup>2</sup>

This paper presents results of an empirical study on fitting log-linear models to data on estimates of characteristics and their coefficients of variation (CV) from the Canadian Labour Force Survey. The characteristics were classified into groups on the basis of design effects and models were fitted to data on estimates of characteristic totals and their CVs over twelve month period. The models can be used in situations where estimates of CV are needed for new characteristics, and for providing more precise estimates of reliability of estimates based on past data. The problem of evaluation of fit of the models is considered.

## 1. INTRODUCTION

This paper presents results of an evaluation study on models for estimation of coefficient of variation (CV) of estimates of characteristics based on the Canadian Labour Force Survey (LFS). The LFS is a monthly household survey with a stratified multi-stage area sample design with a sample size of approximately 55,000 households.

Each month estimates of CV are calculated for a set of characteristics using Keyfitz method of variance estimation based on Taylor series approximation [4], [5]. However, computation of appropriate variance estimates for all estimates tabulated from a large scale survey such as the LFS is not possible due to operational constraints of time and

---

<sup>1</sup> Presented at the American Statistical Association Annual Meeting in Detroit, August 1981.

<sup>2</sup> P. D. Ghangurde, Census and Household Survey Methods Division, Statistics Canada.

costs. The model-based estimates of CV can be used to obtain preliminary estimates of reliability for new characteristics based on the past data, and when estimates of CV for an extended period (e.g. one year) are needed. The models can also be used for obtaining concise estimates of reliability, e.g. alphabetic indicators for ranges of CV.

In section 2 the linear and non-linear models used for estimation of totals and proportions are explained. Sections 3 and 4 review considerations made in forming groups, fitting models and evaluation of goodness of fits.

## 2. THE MODELS

The LFS is a monthly household survey in which dwelling is the final stage sampling unit. Each of the ten provinces in Canada are divided into economic regions which consist of groups of counties with similar economic structure. The economic regions are divided into geographic strata and multi-stage area samples are drawn without replacement with two stages in self-representing strata in the large urban centres and three or four stages in the non-self-representing strata in rural areas. The sample selection in the initial stages is with probability proportional to population size and that in the last stage, in which dwellings are selected from clusters, being systematic.

The design-based estimates within strata are obtained by weighting the data by inverse of probabilities of selection. An adjustment of the basic weight for non-response and ratio estimation within age-sex groups, which are post-strata, is used to obtain final estimates. The census-based population projections for age-sex groups within each province are used as auxiliary variable totals for ratio estimation. More details on the sample design and estimation are given in [5].



The variance estimates of various characteristics at the province level are obtained by Taylor series approximation assuming that the primary sampling units (psus) within non-self-representing strata are selected independently. In self-representing strata the sampled clusters are divided into two groups, which are treated as pseudo-psus and are assumed to have been selected independently. The variance estimate for an estimated characteristic total at Canada level is the sum of corresponding provincial variance estimates [5]. The variance of an estimate  $\hat{X}$  of a characteristic total  $X$  in a province can also be expressed as

$$V(\hat{X}) = F (W-1) X (1 - \frac{X}{P}), \quad (1)$$

where  $P$  = population for the province,

$W$  = inverse sampling ratio,

$F$  = design effect for the characteristic, and

$n$  = sample size (persons).

The expression (1) for  $V(\hat{X})$  relates the variance obtained for the complex ratio estimate based on a stratified multi-stage sample design to the variance of the estimate based on a simple random sample of the same size drawn from the finite population of size  $P$ . The sampling variance of an estimate of total based on a simple random sample of size  $n$  ( $= \frac{P}{W}$ ) is the usual binomial variance with finite population correction. The term,  $F$ , the design effect, represents a factor by which variance is increased due to the effect of such factors as sampling procedure at each stage, the extent of stratification and post-stratification, size of units at various stages and clustering of counts of the characteristic in the province. It may be noted that stratification and post-stratification usually reduce the variance and clustering increases variance of an estimate.

In general, design effects tend to be greater than one due to clustered sample design of the LFS. The labour force status categories such as "employed", "unemployed" by age-sex groups tend to have lower design effects due to post-stratification by age-sex which decreases their variance. Those for labour force status by particular industry tend to

be large due to their location in specific areas. Design effects are known to be related to measures of homogeneity and average size of clusters. Models expressing their relationships have been developed for many surveys. In a study on components of variance in the LFS the design effects and measures of homogeneity have been analyzed for a number of characteristics [2].

A measure of precision of estimates which is independent of the level of the estimate and the scale is coefficient of variation. The  $CV(\hat{X})$  is given by

$$CV(\hat{X}) = \sqrt{F(W-1) \left( \frac{1}{X} - \frac{1}{P} \right)} \quad (2)$$

By taking logarithms to base e on both sides of (2) we have an equation relating CV, X and P given by

$$\log CV(\hat{X}) = \frac{1}{2} \log F(W-1) - \frac{1}{2} \log X + \frac{1}{2} \log \left( 1 - \frac{X}{P} \right). \quad (3)$$

Because of the third term on the right, the equation (3) is not linear in  $\log CV$  and  $\log X$ , even if  $F(W-1)$  is assumed constant. However, for small values of  $X$  the contribution of the third term is negligible. A model based on (3) is given by

$$\log CV(\hat{X}) = A + B \log X + \epsilon, \quad (4)$$

where A and B are parameters of the model and  $\epsilon$  is the error term. The estimate of parameter B will differ from  $-\frac{1}{2}$  depending on the extent to which  $B \log X$  approximates  $\frac{1}{2} \log \left[ X / \left( 1 - \frac{X}{P} \right) \right]$  over the range of X. In an evaluation of fits of (4) and of an alternative model (5) given by

$$\log CV(\hat{X}) = A + B \log \frac{X}{\left( 1 - \frac{X}{P} \right)} + \epsilon, \quad (5)$$

the goodness of fit for the two models as shown by  $R^2$ , the ratio of regression sum of squares to total sum of squares, was found to be

quite close. The model (4) is linear in  $\log X$  and  $\log CV$  and is simpler than model (5).

A non-linear model corresponding to (4) is given by:

$$CV(\hat{X}) = A' X^{B'} + \epsilon, \quad (6)$$

where  $A'$  and  $B'$  are parameters of the model and  $\epsilon$  is the error term. The two models (4) and (6) were fitted to data on monthly estimates and their CVs for 90 characteristics in each of 10 provinces and Canada.

### 3. GROUPING OF CHARACTERISTICS

The monthly design effects of LFS estimates for January-December 1980 for each of 90 characteristics excluding total population for each province and Canada were averaged and plotted to decide the ranges for the two groups. In each province, the first group consists of characteristics with design effects greater than  $D$ .

Table 1 shows the boundary values  $D$  for group I and II in each province and at Canada level, and the number of characteristics in group II. The grouping of characteristics was done by arranging characteristics in increasing order of average design effects. The boundary value  $D$  was selected so that the assumption of equal design effects was satisfied as far as possible in group I. The second group consists of all remaining characteristics where the assumption of equal design effects is more crude. Most characteristics pertaining to labour force status by age-sex groups fall in group I. "Employed by industry" and "duration of unemployment" mostly fall in group II. The average design effects differ substantially between provinces and for Canada. More refined grouping of characteristics on the basis of models for design effects is being investigated.

It may be noted that about 80% of the characteristics in each province and for Canada, have been classified in group I. For obtaining a

conservative estimate of CV for a new characteristic models based on group II can be used. For a characteristic for which monthly estimates of CV are routinely produced the models for the group in which the characteristic falls, can be used to obtain approximate estimate of CV with a greater precision than that based on monthly data.

In the following section the assumptions made in fitting the models (4) and (6) are explained and model fits are evaluated.

#### 4. EVALUATION OF MODELS

The basis of fitting the log-linear model (4) is to treat the model as a simple linear regression model in  $y = \log CV(\hat{X})$  and  $x = \log X$  and to obtain estimates of parameters A and B in the linear regression framework. The usual assumptions of independence of errors and constant variance have been made. Under these assumptions,  $R^2$  provides a measure of fit of the model. The values of the estimated parameters and coefficients of determination,  $R^2$ , for group I and II in 10 provinces and Canada are given in Table 2. The actual fitting of these models was done by using SAS utility.

All  $R^2$  values are significant and quite high indicating that the fits are very good. The error plots do not show any patterns to conclude that the assumption of constant variance is not satisfied. Under these assumptions and normality of errors  $CV(\hat{X})$  has a log-normal distribution with constant CV for any value of X.

The non-linear model (6) was fitted by Gauss-Newton method using SAS utility. The initial values of parameters  $A'$  and  $B'$  were assumed to be 1.00 and -0.50 respectively. The number of iterations required to reach convergence was at most 8 for each province and Canada, the convergence criterion being that the relative difference between successive error sum of squares is less than  $10^{-8}$ . Table 3 shows values of estimated parameters and errors sum of squares for Canada Group II. The errors are approximately normally distributed as shown by normal probability plots.

Since it is of interest to compare the fits of the non-linear model for provinces, Canada and the two groups it is necessary to have a criterion of goodness of fit. In the non-linear model, the total sum of squares is not equal to the total of regression and error sums of squares. A criterion  $R'^2$  can be defined as

$$R'^2 = 1 - \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2},$$

where  $\hat{Y}_i$ 's are estimated CVs based on the model,  $Y_i$ 's are observed CVs and  $\bar{Y}$  their mean. The summation extends over N, the number of characteristics in the group multiplied by 12, the number of months. In the linear case  $R^2 = R'^2$ . However, in the non-linear case  $R^2 \neq R'^2$  since the total sum of squares is not equal to regression sum of squares plus error sum of squares due to product term not being zero.

The errors  $(Y_i - \hat{Y}_i)$  will be small when the fit is good giving a value of  $R'^2$  close to 1, the errors  $(Y_i - \hat{Y}_i)$  will be large when the fit is poor giving a small value of  $R'^2$ . When all the points lie on the fitted curve i.e.  $Y_i = \hat{Y}_i$  for all i,  $R'^2 = 1$ . However, in general no lower bound to  $R'^2$  seems to exist. The values of  $R'^2$  shown in Table 4 tend to be greater for group I as compared to group II, which has 13 to 21 characteristics out of the total of 90.

Although the log-linear model (4) was fitted to data on logarithms of estimates and their CVs and its fit seems to be good, the fitted models for provinces and Canada are used for estimation of CV of estimates. In order to compare the fit of the transformed model to original data of estimates and their CVs, these data and the transformed model corresponding to (4) were plotted for the two groups in 10 provinces and Canada. From these charts it can be concluded that the transformed model corresponding to (4) fits the data of estimates and their CVs better than the non-linear



model (6), especially for small values of estimates. The plots of these models for Canada group 11 are shown on Chart 1 and 2.

## 5. CONCLUDING REMARKS

The characteristics considered are total persons with labour force status by age-sex, industry, marital status and total persons with various ranges of duration of unemployment. However, the models can also be used for proportions instead of totals. The models are not applicable to estimates for subprovincial areas such as urban centres or groups of economic regions, since design effects for these areas are more unstable and can be much higher due to the effect of ratio-adjustment based on projected population at province level [1].

An assumption made in the use of models for a new characteristic is that its design effect is close to the average for the group. This requires finer grouping of characteristics of various types possibly on the basis of models relating design effects with measures of homogeneity for these characteristics. In fitting the models, it was assumed that errors are uncorrelated and that independent variable is fixed. Since twelve monthly estimates for each characteristic were used, there could be correlation in errors for estimates for a given characteristic. Extension of the study to models with errors in independent variable and correlated errors is being considered.

A problem in evaluation of fit of non-linear models, whether actually fitted to data or transformed from linear models, is the lack of a criterion for comparison of fits of different models. The criterion suggested in section 4 may be appropriate for comparison of fits of a model to different data sets, but may not work for different models.

TABLE 1: DESIGN EFFECT BOUNDARY VALUES AND NUMBERS OF CHARACTERISTICS  
IN GROUPS I AND II\*

Province	Boundary Value (D)	Number of Characteristics	
		Group I	Group II
Newfoundland	2.3	75	15
P.E.I.	1.9	73	17
Nova Scotia	1.9	74	16
New Brunswick	2.2	77	13
Quebec	1.9	73	17
Ontario	1.7	69	21
Manitoba	2.0	76	14
Saskatchewan	2.8	76	14
Alberta	2.1	71	19
British Columbia	2.3	73	17
Canada	1.9	77	13

\* A characteristic belongs to Group I if its design effect (averaged over the 12-month period from January to December 1980) is less than or equal to the boundary value D. If the average design effect is greater than D, then the characteristics is in Group II.

TABLE 2: REGRESSION COEFFICIENTS AND  $R^2$  FOR LOG-LINEAR MODEL

Province	Group	Regression Coefficient		$R^2$
		A	B	
Newfoundland	I	3.3119	-0.5723	0.9534
	II	3.7757	-0.6101	0.9377
P.E.I.	I	2.7962	-0.5617	0.9485
	II	3.1796	-0.5885	0.8887
Nova Scotia	I	3.4612	-0.5837	0.9702
	II	3.6412	0.5257	0.8717
New Brunswick	I	3.2782	-0.5545	0.9606
	II	3.7544	-0.6017	0.9357
Quebec	I	4.3298	-0.5942	0.9686
	II	4.3093	-0.5216	0.9127
Ontario	I	4.3825	-0.6053	0.9736
	II	4.1796	-0.5009	0.9633
Manitoba	I	3.5155	-0.5926	0.9619
	II	3.8769	-0.5640	0.9166
Saskatchewan	I	3.3796	-0.5700	0.9544
	II	3.5478	-0.4423	0.8994
Alberta	I	3.6960	-0.5968	0.9678
	II	3.7526	-0.5090	0.9513
B.C.	I	3.9847	-0.5750	0.9621
	II	3.9814	-0.4708	0.8410
Canada	I	4.3458	-0.5936	0.9703
	II	4.2357	-0.5191	0.9699

TABLE 3: NON-LINEAR LEAST SQUARES: GAUSS-NEWTON METHOD

CANADA (GROUP II)

Iteration	A'	B'	Residual S.S.
0	1.00000000	-0.50000000	3401.93232121
1	15.22076853	-0.23647629	461.76322678
2	26.47981387	-0.36743343	322.67707190
3	51.94184546	-0.51147529	248.68405130
4	57.29455529	-0.47434886	99.32440727
5	58.32558100	-0.48419609	96.57832290
6	58.28627964	-0.48409502	96.57810754
7	58.28746710	-0.48409960	96.57810746

TABLE 4:  $R^2$  FOR GROUP I AND II

Province	Group	N*	$R^2 = 1 - \frac{\text{Error S.S.}}{\text{Total S.S.}}$
Newfoundland	I	866	0.9362
	II	190	0.8835
P.E.I	I	827	0.8925
	II	294	0.7285
Nova Scotia	I	872	0.9790
	II	192	0.7813
New Brunswick	I	908	0.9990
	II	156	0.8639
Quebec	I	859	0.9800
	II	204	0.7804
Ontario	I	823	0.9632
	II	252	0.9208
Manitoba	I	895	0.9691
	II	168	0.8137
Saskatchewan	I	896	0.9436
	II	168	0.8196
Alberta	I	845	0.9701
	II	228	0.8852
B.C.	I	868	0.9319
	II	204	0.7786
Canada	I	923	0.9665
	II	156	0.9286

\* N for group I can be less than 12 (no. of characteristics) due to exclusion of characteristics with zero estimates.



SYMBOL USED IS \*  
LEGEND: A = 1 OBS, B = 2 OBS, ETC.

PLOT OF CVHAT1\*EST  
PLOT OF CV\*EST



NOTE: 98 OBS HIDDEN

# CHART 2

#3 SIMPLE POWER MODEL  
PROV=CANADA GROUP=2

PLOT OF PRE3\*EST SYMBOL USED IS \*  
PLOT OF CV\*EST LEGEND: A = 1 OBS, B = 2 OBS, ETC.



\*\*\*\*\*  
ACBEA A ACC AB  
\*\*\*\*\*  
BCCABA

#### ACKNOWLEDGEMENTS

The author would like to thank the referee for helpful comments.

#### REFERENCES

- [1] Ghangurde, P.D. and Gray, G.B. (1981), "Estimation for Small Areas in Household Surveys", Communications in Statistics, Theory and Methods, A 10(22), 2327-38.
- [2] Gray, G.B. and Platek, R. (1976), "Analysis of Design Effects and Variance Components in Multistage Surveys", Survey Methodology, Vol. 2, No. 1., 1-30.
- [3] Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors", Presented at Meeting of the International Association of Survey Statisticians.
- [4] Keyfitz, N. (1957), "Estimates of Sampling Variance Where Two Units are selected from Each Stratum". Journal of the American Statistical Association, 52, 503-510.
- [5] Platek, R. and Singh, M.P. (1976), Methodology of the Canadian Labour Force Survey. Catalogue 71-526 occasional.
- [6] Sprent, P. (1969) "Models in Regression and Related Topics", Methuen and Co.

The Editorial Board wish to thank Mr. R.E. Drover, Chairman, Publication Board and the staff of Administrative Services Division for their continuing support of the production of this Journal.

Acknowledgement is also due to N. Brien, M. Fluet, P. Foy and G. Kriger for their proofreading and preparation of final content.

Thanks are also due to Mrs. D. Edirisinghe, for her patient typing of the Journal, as well as for the execution of numerous other duties associated with its production. Finally, the Editorial Board wish to thank the following persons who have served as referees during the past year.

D.A. Binder	G. Kriger
R.G. Carter	S. Kumar
G.H. Choudhry	M.L. Lawes
D.P. Dixon	I. Macredie
J.D. Drew	M.J. March
S. Earwaker	A. Satin
M. Fluet	K.P. Srinath
P. Foy	L. Swain
G.B. Gray	P.F. Timmons
M.A. Hidioglou	

# SURVEY METHODOLOGY

June 1981

Vol. 7

No. 1

A Journal produced by Methodology Staff, Statistics Canada

## C O N T E N T S

Survey Maintenance - Philosophy and Practice F. MAYDA and P. TIMMONS .....	1
Imputation in Surveys: Coping with Reality I.G. SANDE .....	21
Redesigning Continuous Surveys in a Changing Environment M.P. SINGH and J.D. DREW .....	44
For-hire Trucking Survey: Survey Design R. LUSSIER .....	74
Construction of Working Probabilities and Joint Selection Probabilities for Fellegi's PPS Sampling Scheme G.H. CHOUDHRY .....	93







Préparé par les méthodologistes de Statistique Canada

TABLE DES MATIÈRES

Coordination des enquêtes - philosophie et pratique	1
F. MAYDA et P. TIMMONS	.....
Imputation dans les enquêtes: affrontes la réalité	21
J.G. SANDE	.....
Le ramaniement des enquêtes permanentes dans un milieu en mutation	44
M.P. SINGH et J.D. DREW	.....
La méthodologie de l'enquête sur le transport routier de marchandises pour le compte d'autrui	74
R. LUSSIER	.....
Calcul des "probabilités de travail" et des probabilités conjointes de sélection de la méthode d'échantillonnage à probabilités inégales de Felllegi	93
G.H. CHOUDHRY	.....

Le comité de rédaction désire remercier M. R.E. Drover, Président, Comité de publication et le personnel de la Division des services administratifs pour leur appui constant dans la publication de cette revue.

De même que N. Brien, M. Fluet, P. Foy et G. Kriger pour leur aide lors de la révision des textes et la préparation de la version finale.

Nous remercions également Mme. D. Edirisinghe, qui a dactylographié soigneusement la revue, et qui s'est acquittée de nombreuses autres tâches associées à sa publication.

Le comité de rédaction désire enfin remercier les personnes suivantes, qui ont bien voulu faire la critique des articles présentés au cours de l'année dernière.

D.A. Binder	G. Kriger
R.G. Carter	S. Kumar
G.H. Choudhry	M.L. Lawes
D.P. Dixon	I. Macredie
J.D. Drew	M.J. March
S. Earwaker	A. Satin
M. Fluet	K.P. Srinath
P. Foy	L. Swain
G.B. Gray	P.F. Timmons
M.A. Hidiroglou	

# REMERCIEMENTS

L'auteur aimerait remercier l'arbitre pour ses commentaires et ses observations.

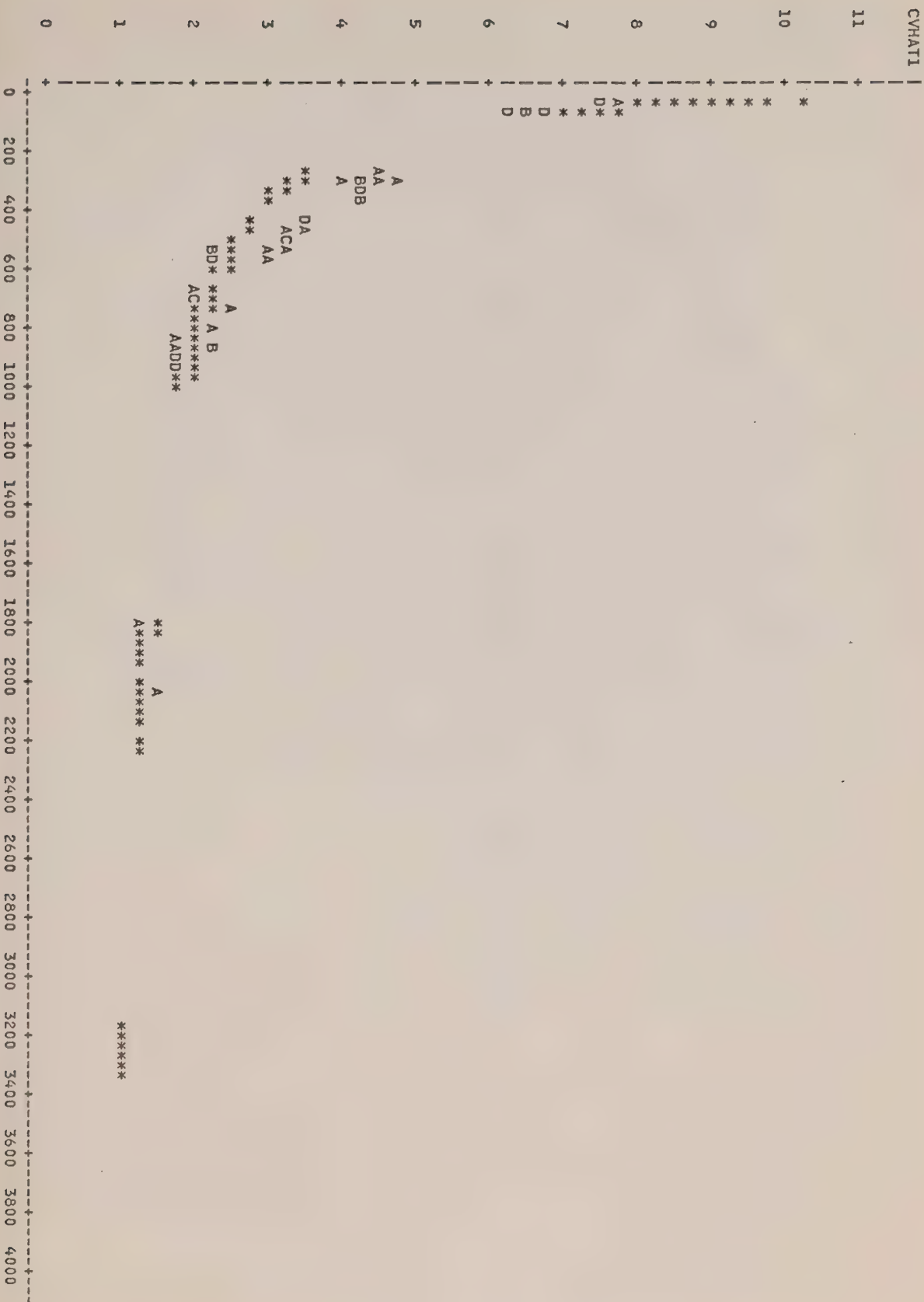
## BIBLIOGRAPHIE

- [1] Ghangurde, P.D. et Gray, G.B. (1981), "Estimation for Small Areas in Household Surveys", Communications in Statistics Theory and Methods, A 10(22), 2327-38.
- [2] Gray, G.B. et Platek, R. (1976), "Analysis of Design Effects and Variance Components in Multistage Surveys", Techniques d'enquête, vol. 2, n° 1, 1-30.
- [3] Kalton, G. (1977), "Practical Methods for Estimating Survey Sampling Errors", exposé présenté à une assemblée de l'International Association of Survey Statisticians.
- [4] Keyfitz, N. (1957), "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum", Journal of the American Statistical Association, 52, 503-510.
- [5] Platek, R. et Singh, M.P. (1976), Méthodologie de l'enquête sur la population active, n° 71-526 au catalogue, hors-série.
- [6] Sprent, P. (1969) "Models in Regression and Related Topics", Methuen and Co.

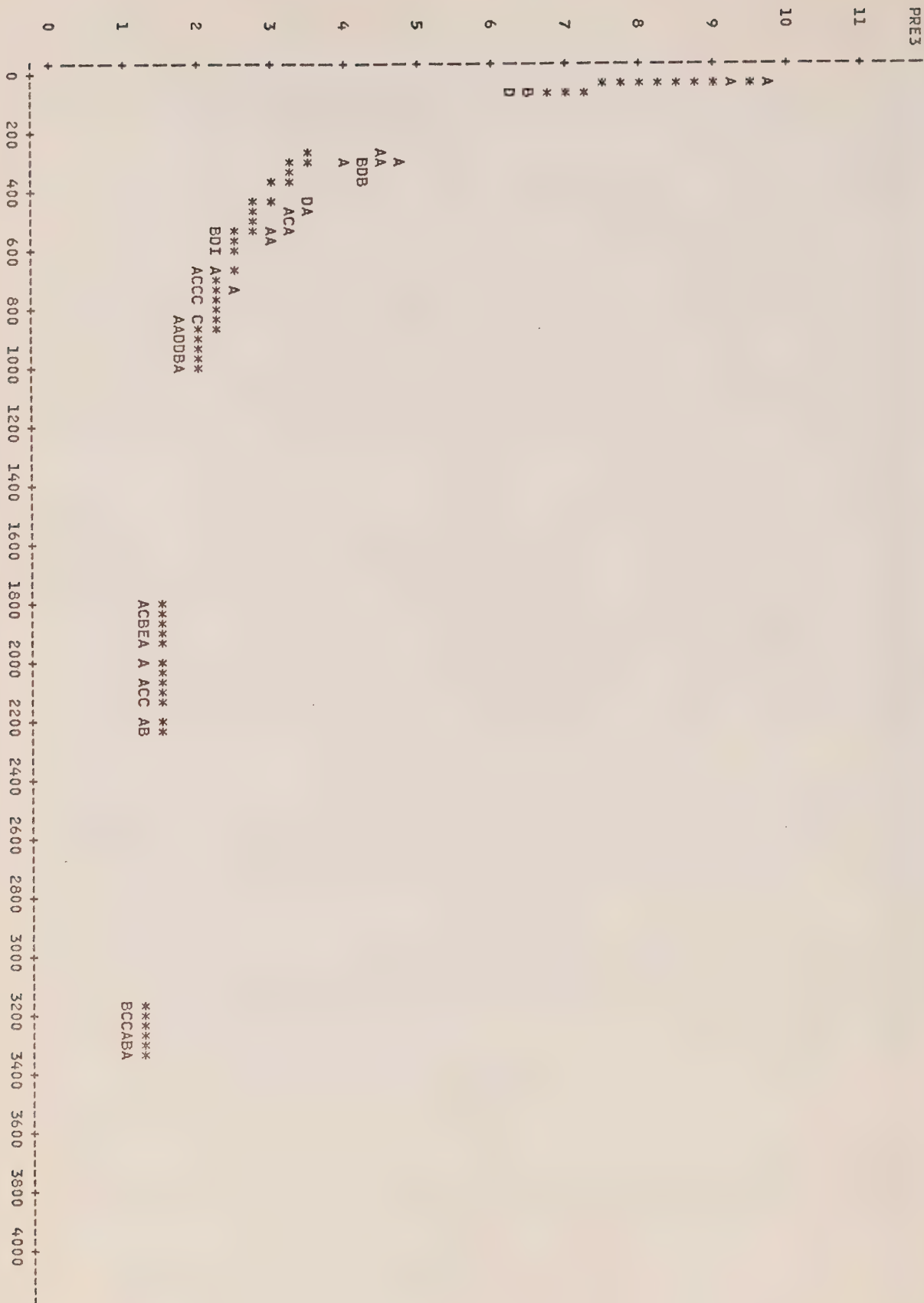


PLOT OF CVHAT1\*EST      SYMBOL USED IS \*

PLOT OF CV\*EST      LEGEND: A = 1 OBS, B = 2. OBS, ETC.



EST



NOTE: 101 OBS HIDDEN

EST

TABLEAU 4: R<sup>2</sup> DANS LES GROUPES I ET II

Province      Groupe      N\*      R<sup>2</sup> = 1 -  $\frac{\text{Somme des carrés résiduelle}}{\text{Somme des carrés totale}}$

Terre-Neuve	I	866	0.9362	0.8835
Ile-du-Prince-Édouard	I	827	0.8925	0.7285
Nouvelle-Écosse	I	872	0.9790	0.7813
Nouveau-Brunswick	I	908	0.9990	0.8639
Québec	I	859	0.9800	0.7804
Ontario	I	823	0.9632	0.9208
Manitoba	I	895	0.9691	0.8137
Saskatchewan	I	896	0.9436	0.8196
Alberta	I	845	0.9701	0.8852
C.-B.	I	868	0.9319	0.7786
Canada	I	923	0.9665	0.9286
	II	156		

\* N dans le groupe I peut être inférieur à 12 (nombre de caractéristiques) à cause de l'exclusion des caractéristiques sans estimation.

TABEAU 3: MOINDRES CARRÉS NON LINÉAIRES: MÉTHODE DE GAUSS-NEWTON

CANADA (GROUPE II)

Itération	A'	B'	Somme des carrés résiduelle
0	1.00000000	-0.50000000	3401.93232121
1	15.22076853	-0.23647629	461.76322678
2	26.47981387	-0.36743343	322.67707190
3	51.94184546	-0.51147529	248.68405130
4	57.29455529	-0.47434886	99.32440727
5	58.32558100	-0.48419609	96.57832290
6	58.28627964	-0.48409502	96.57810754
7	58.28746710	-0.48409960	96.57810746

TABLEAU 2: COEFFICIENTS DE RÉGRESSION ET R<sup>2</sup> SELON LE MODÈLE LINÉAIRE LOGARITHMIQUE

Province	Groupe	A	B	R <sup>2</sup>
Terre-Neuve	I	3.3119	-0.5723	0.9534
	II	3.7757	-0.6101	0.9377
Ile-du-Prince-Édouard	I	2.7962	-0.5617	0.9485
	II	3.1796	-0.5885	0.8887
Nouvelle-Écosse	I	3.4612	-0.5837	0.9702
	II	3.6412	0.5257	0.8717
Nouveau-Brunswick	I	3.2782	-0.5545	0.9606
	II	3.7544	-0.6017	0.9357
Québec	I	4.3298	-0.5942	0.9686
	II	4.3093	-0.5216	0.9127
Ontario	I	4.3825	-0.6053	0.9736
	II	4.1796	-0.5009	0.9633
Manitoba	I	3.5155	-0.5926	0.9619
	II	3.8769	-0.5640	0.9166
Saskatchewan	I	3.3796	-0.5700	0.9544
	II	3.5478	-0.4423	0.8994
Alberta	I	3.6960	-0.5968	0.9678
	II	3.7526	-0.5090	0.9513
C.-B.	I	3.9847	-0.5750	0.9621
	II	3.9814	-0.4708	0.8410
Canada	I	4.3458	-0.5936	0.9703
	II	4.2357	-0.5191	0.9699



TABLEAU 1: VALEURS LIMITES DE L'EFFET DU PLAN D'ÉCHANTILLONNAGE ET NOMBRE DE CARACTÉRISTIQUES DANS LES GROUPES I ET II\*

Province	Valeur limite (D)	Nombre de caractéristiques	Groupe I	Groupe II
Terre-Neuve	2.3	75	15	
Ile-du-Prince-Édouard	1.9	73	17	
Nouvelle-Écosse	1.9	74	16	
Nouveau-Brunswick	2.2	77	13	
Québec	1.9	73	17	
Ontario	1.7	69	21	
Manitoba	2.0	76	14	
Saskatchewan	2.8	76	14	
Alberta	2.1	71	19	
Colombie-Britannique	2.3	73	17	
Canada	1.9	77	13	

\* Une caractéristique est classée dans le groupe I si l'effet du plan d'échantillonnage pour cette caractéristique (dont la moyenne est calculée sur une période de 12 mois, soit de janvier à décembre 1980) est inférieur ou égal à la valeur limite D. Si l'effet moyen est supérieur à D, la caractéristique est alors incluse dans le groupe II.

diverses séries de données, mais il est probable également que cela ne convienne pas à des modèles différents.

Le cas des petites valeurs. Les graphiques 1 et 2 illustrent ces modèles appliqués aux valeurs du groupe II à l'échelle du Canada.

## 5. CONCLUSION

Les caractéristiques étudiées sont le nombre total de personnes classées selon leur situation vis-à-vis de l'activité par groupe d'âge/sex, secteur d'activité et état matrimonial, de même que le nombre total de personnes selon la durée de la période de chômage. Les modèles peuvent être utilisés aussi bien pour la production de proportions que pour celle des totaux, mais ils ne peuvent pas être appliqués aux estimations relatives aux régions infraprovinciales, notamment les centres urbains ou les groupes de régions économiques, parce que les effets du plan d'échantillonnage dans ces régions sont moins stables et peuvent être beaucoup plus élevés à cause de l'effet de la correction du quotient effectuée à l'aide des projections démographiques provinciales [1].

On a posé comme hypothèse d'application de modèles pour une nouvelle caractéristique que l'effet du plan d'échantillonnage est proche de la moyenne établie pour le groupe. Il faut alors détailler encore plus le groupement des caractéristiques en fonction de modèles qui permettent de lier les effets du plan aux mesures d'homogénéité de ces caractéristiques. Pour l'ajustement des modèles, on a supposé que les erreurs ne sont pas corrélées et que la variable indépendante est fixe. Comme on a utilisé 12 estimations mensuelles de chaque caractéristique, il est possible que les erreurs relatives aux estimations d'une caractéristique particulière soient corrélées. On examine actuellement la possibilité d'étendre cette étude de façon à englober les modèles qui tiennent compte des erreurs produites avec la variable indépendante et des erreurs corrélées.

Le problème qui se pose quant à l'évaluation de l'ajustement de modèles non linéaires, qu'ils soient effectivement ajustés aux données ou qu'il s'agisse de modèles linéaires transformés, est l'absence d'une norme de comparaison des ajustements de divers modèles. Il est probable que la norme proposée à la partie 4 convienne pour la comparaison des ajustements de modèles à

La régression et de la somme des carrés résiduelle. La norme  $R^2$  peut être exprimée

$$R^2 = 1 - \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}$$

où les  $\hat{Y}_i$  sont les CV estimés basés sur le modèle, les  $\bar{Y}$  sont les CV observés et  $\bar{Y}$  leur moyenne. La sommation s'étend à N, qui est le nombre de caractéristiques dans le groupe, multiplié par 12, le nombre de mois. Dans le modèle linéaire,  $R^2 = R^2$ . Cependant, dans le modèle non linéaire,  $R^2$  est différent de  $R^2$ , car la somme des carrés totale n'est pas égale à la somme des carrés de la régression plus la somme des carrés résiduelle parce que le produit n'est pas zéro.

Les erreurs ( $Y_i - \hat{Y}_i$ ) sont petites lorsque l'ajustement est précis et produit une valeur de  $R^2$  proche de 1; les erreurs ( $Y_i - \hat{Y}_i$ ) sont importantes lorsque l'ajustement est faible et produit une petite valeur de  $R^2$ . Lorsque tous les points se trouvent sur la courbe ajustée, c'est-à-dire que  $Y_i = \hat{Y}_i$  pour tous les i, alors  $R^2 = 1$ . Toutefois, règle générale, il ne semble pas y avoir de borne inférieure à  $R^2$ . Les valeurs de  $R^2$  présentées dans le tableau 4 semblent être plus élevées dans le groupe I que dans le groupe II qui comprend de 13 à 21 caractéristiques sur 90.

Bien que le modèle linéaire logarithmique (4) ait été ajusté en fonction des logarithmes des estimations et de leur CV et que l'ajustement semble précis, les modèles ajustés pour les provinces et le Canada sont appliqués à l'estimation des CV des estimations. Afin de pouvoir comparer l'ajustement du modèle transformé aux estimations initiales et à leur CV, les données et le nouveau modèle (4) ont été portés sur un graphique, et ce pour les deux groupes relatifs aux 10 provinces et au Canada. À partir de ces diagrammes, on peut conclure que ce nouveau modèle (4) peut ajuster les estimations et leur CV de façon plus précise que le modèle non linéaire (6), surtout dans

#### 4. EVALUATION DES MODELES

Pour ajuster le modèle linéaire logarithmique (4), il faut essentiellement considérer ce modèle comme un modèle de régression linéaire simple sous la forme  $y = \log CV(\hat{X})$  et  $x = \log X$ , et calculer les estimations des paramètres A et B selon la base de régression linéaire. Les hypothèses usuelles d'indépendance des erreurs et de variance constante s'appliquent et, selon ces hypothèses,  $R^2$  fournit une mesure de l'ajustement du modèle. Les valeurs des paramètres estimés et les coefficients de détermination,  $R^2$ , dans les groupes I et II, relativement aux 10 provinces et au Canada, figurent au tableau 2. L'ajustement effectif de ces modèles a été fait à l'aide du programme SAS.

Toutes les valeurs  $R^2$  sont significatives et assez élevées, ce qui démontre que les ajustements sont très précis. Le graphique des erreurs ne révèle pas de tendance qui permette de conclure que l'hypothèse de la variance constante n'est pas satisfaite. En vertu de ces hypothèses et compte tenu de la normalité des erreurs,  $CV(\hat{X})$  présente une distribution log-normale avec un CV constant pour n'importe quelle valeur de X.

Le modèle non linéaire (6) a été ajusté selon la méthode Gauss-Newton à l'aide du programme SAS. On a supposé que les valeurs initiales des paramètres A' et B' étaient 1.00 et -0.50 respectivement. Le nombre d'itérations requises pour atteindre la convergence a été de 8 au maximum pour chaque province et pour le Canada, la norme de convergence étant que l'écart relatif entre plusieurs sommes des carrés résiduelles doit être inférieur à 10<sup>-8</sup>. Le tableau 3 contient les valeurs des paramètres estimés et la somme des carrés résiduelle pour le Canada, dans le groupe II. Les erreurs affichent une distribution approximativement normale puisque les points sur le graphique révèlent une distribution normale de probabilités.

Comme la comparaison des ajustements du modèle non linéaire appliqué aux provinces, au Canada, pour les deux groupes, présente un certain intérêt, il est nécessaire de définir une norme de validité de l'ajustement. Dans le modèle non linéaire, la somme des carrés totale n'est pas égale au total de



des deux groupes. Dans chaque province, le premier groupe comprenait les caractéristiques affichant des effets plus grands que D.

Le tableau 1 indique les valeurs limites D des groupes I et II dans chaque province et pour le Canada, de même que le nombre de caractéristiques incluses dans le groupe II. Pour grouper les caractéristiques, on les a classées par ordre croissant de la moyenne des effets du plan. La valeur limite D a été choisie de façon que l'hypothèse de l'égalité des effets du plan soit satisfaite le plus possible dans le groupe I. Le deuxième groupe comprend toutes les autres caractéristiques pour lesquelles l'égalité des effets du plan est supposée plus rudimentaire. La plupart des caractéristiques relatives à la situation vis-à-vis de l'activité, classées par groupe âge/sexe se retrouvent dans le groupe I. Les personnes occupées par secteur d'activité et la durée de la période de chômage entrent habituellement dans le groupe II. Les effets moyens varient beaucoup entre les provinces et pour le Canada. On examine actuellement l'application d'un groupement des caractéristiques plus détaillé basé sur des modèles des effets du plan.

Il convient de souligner qu'environ 80 % des caractéristiques dans chaque province et pour le Canada ont été classées dans le groupe I. Pour produire une estimation prudente du CV de nouvelles caractéristiques, on peut utiliser des modèles basés sur le groupe II. Dans le cas d'une caractéristique dont les estimations mensuelles du CV sont calculées régulièrement, on peut appliquer les modèles construits pour le groupe dans lequel cette caractéristique est classée, si on veut obtenir une estimation approximative du CV qui sera plus précise que celle basée sur les données mensuelles.

Dans la partie suivante, nous examinons les hypothèses appliquées à l'ajustement des modèles (4) et (6) et nous évaluons ces ajustements.

Toutefois, dans le cas des petites valeurs de  $X$ , le troisième terme a un effet négligeable. Un modèle fondé sur (3) est exprimé par

$$(4) \quad \log CV(X) = A + B \log X + E,$$

où  $A$  et  $B$  sont les paramètres du modèle et  $E$  le terme de l'erreur. L'estimation du paramètre  $B$  sera différente de  $-\frac{1}{2}$  selon que  $B \log X$  est plus ou moins une approximation de  $\frac{1}{2} \log [X/(1 - \frac{p}{X})]$  dans l'étendue de  $X$ . Dans une évaluation des ajustements du modèle (4) et du modèle de rechange (5) donné par

$$(5) \quad \log CV(X) = A + B \log \frac{X}{(1 - \frac{p}{X})} + E,$$

on a constaté que la validité de l'ajustement, exprimée par  $R^2$ , soit le rapport de la somme des carrés de la régression sur la somme des carrés totale, était sensiblement la même pour les deux modèles. Le modèle (4) est linéaire sous la forme  $\log X$  et  $\log CV$  et peut être appliqué plus facilement que le modèle (5).

Un modèle non linéaire correspondant à (4) prend la forme:

$$(6) \quad CV(X) = A'X^{B'} + E,$$

où  $A'$  et  $B'$  sont les paramètres du modèle et  $e$  est le terme de l'erreur. Les deux modèles (4) et (6) ont été ajustés en fonction des estimations mensuelles et du (CV) de 90 caractéristiques dans chacune des 10 provinces.

### 3. GROUPEMENT DES CARACTÉRISTIQUES

On a calculé la moyenne des effets du plan mensuels des estimations de l'EPA pour la période de janvier à décembre 1980 et pour chacune des 90 caractéristiques, sauf la population totale de chaque province et du Canada, puis on les a portées sur un graphique afin de déterminer l'étendue

constitution de grappes à partir des totaux obtenus pour la caractéristique dans la province. On doit souligner ici que la stratification et la stratification à posteriori ont habituellement pour effet de réduire la variance d'une estimation, alors que la constitution de grappes l'augmente.

De façon générale, les effets du plan d'échantillonnage tendent à être plus importants que les effets dus à l'échantillonnage par grappes des unités de l'EPA. Les effets du plan tendent à être moins élevés dans le cas des catégories de la population active comme celles des "personnes occupées" et des "chômeurs" par groupe âge/sexe, en raison de la stratification à posteriori par âge et sexe qui réduit leur variance. Lorsque la population active est classée selon le secteur d'activité, les effets du plan tendent à être plus élevés à cause de la concentration dans des régions précises. On sait que les effets du plan d'échantillonnage sont fonction des mesures d'homogénéité et de la taille moyenne des grappes. Des modèles fondés sur de tels rapports ont été conçus pour nombre d'enquêtes. Dans une étude sur les composantes de la variance intervenant dans l'EPA, on a analysé les effets du plan d'échantillonnage et les mesures d'homogénéité d'un certain nombre de caractéristiques [2].

Le coefficient de variation nous donne une mesure de précision des estimations sans égard à leur niveau et à l'échelle. Le  $CV(\hat{X})$  est donné par

$$(2) \quad CV(\hat{X}) = F(W - 1) \left( \frac{X}{1} - \frac{1}{p} \right).$$

En utilisant des logarithmes pour déterminer la base e des deux membres de l'expression (2), nous obtenons une équation reliant CV, X et P qui prend la forme:

$$(3) \quad \log CV(\hat{X}) = \frac{2}{1} \log F(W - 1) - \frac{2}{1} \log X + \frac{1}{2} \log \left( 1 - \frac{p}{X} \right).$$

A cause du troisième terme du deuxième membre, l'équation (3) n'est pas linéaire sous la forme  $\log X$ , même si  $F(W-1)$  est supposé constant.

L'expression (1) pour calculer  $V(\bar{X})$  relie la variance de l'estimation complexe par quotient qui est fondée sur un plan d'échantillonnage stratifié à plusieurs degrés à la variance de l'estimation basée sur un échantillon aléatoire simple de même taille tiré de la population finie de taille P. La variance d'échantillonnage d'une estimation de la valeur totale d'une caractéristique fondée sur un échantillon aléatoire simple de taille  $n(\frac{M}{P})$  correspond à la variance binomiale courante corrigée en fonction de la population finie. Le terme F, l'effet du plan d'échantillonnage, est un facteur d'accroissement de la variance qui réagit à des facteurs comme la méthode d'échantillonnage à chaque degré, le niveau de stratification et de stratification à posteriori, la taille des unités aux divers degrés et la

n = la taille de l'échantillon (nombre de personnes).  
 F = l'effet du plan d'échantillonnage pour la caractéristique

W = l'inverse du taux de sondage

où P = la population de la province

$$V(\bar{X}) = F(W-1) \times (1 - \frac{P}{X}),$$

province donnée prend donc la forme:

variance d'une estimation  $\bar{X}$  du total X d'une caractéristique dans une estimations correspondantes de la variance à l'échelon provincial [5]. La d'une caractéristique à l'échelle nationale consiste à faire la somme des choisies de façon indépendante. L'estimation de la variance du total estimé en deux groupes qui deviennent des pseudo-UPF et sont supposées avoir été les strates autoreprésentatives, les grappes échantillonnées sont réparties strates non autoreprésentatives sont choisies de façon indépendante. Dans supposant que les unités primaires d'échantillonnage (UPF) à l'intérieur des provincial sont obtenues par une approximation de la série de Taylor, en Les estimations de la variance de diverses caractéristiques au niveau

d'échantillonnage et l'estimation, voir la publication [5].  
 dans l'estimation par quotient. Pour plus de détails sur le plan groupe âge/sexes et par province sont les variables auxiliaires utilisées établies à posteriori. Les projections démographiques du recensement par par quotient à l'intérieur des groupes âge/sexes qui sont des strates



Pour produire les estimations relatives aux strates (basées sur le plan d'échantillonnage), on applique aux données un coefficient de pondération qui équivaut à l'inverse des probabilités de sélection. Les estimations finales sont obtenues à la suite d'un ajustement du coefficient de pondération de base pour tenir compte de la non-réponse et par l'estimation systématique.

L'EPA est une enquête mensuelle menée auprès des ménages dans laquelle le logement est l'unité d'échantillonnage du dernier degré. Chaque province du Canada est divisée en régions économiques qui sont formées de groupes de comtés ayant une structure économique semblable. Les régions économiques sont subdivisées en strates géographiques et on tire des échantillons aréolaires sans remise à plusieurs degrés, soit à deux degrés dans les strates autorenseignantes dans les grands centres urbains et à trois ou quatre degrés dans les strates non autorenseignantes dans les régions rurales. Aux premiers degrés, la sélection de l'échantillon se fait selon une probabilité proportionnelle à la taille de la population et, au dernier degré, où les logements sont choisis dans les grappes, la sélection est

## 2. MODELES

Dans la partie 2, on explique les modèles linéaires et non linéaires appliqués au calcul des estimations totales et proportionnelles. Les parties 3 et 4 portent sur les facteurs de groupement des caractéristiques, l'ajustement des modèles et l'évaluation de la validité des ajustements.

Cependant, pour des raisons de temps et d'argent, il est impossible de calculer la variance approximative de toutes les estimations obtenues à partir d'une enquête à grande échelle comme l'EPA. Les estimations des CV basées sur des modèles peuvent être utilisées lorsqu'on veut obtenir des estimations provisoires de la fiabilité de nouvelles caractéristiques fondées sur des données antérieures ou lorsqu'il faut étendre la période d'observation (une année par exemple). Les modèles peuvent également être utilisés pour obtenir des estimations précises de la fiabilité, notamment des indicateurs alphabétiques des intervalles des CV.



MODELES D'ESTIMATION DES ERREURS D'ECHANTILLONNAGE<sup>1</sup>P.D. Ghangurde<sup>2</sup>

Ce document présente les résultats d'une étude empirique sur l'ajustement de modèles linéaires logarithmiques en fonction des estimations de caractéristiques de l'enquête sur la population active et de leurs coefficients de variation (CV). Ces caractéristiques ont été regroupées en fonction des effets du plan d'échantillonnage, et des modèles ont été ajustés en fonction des estimations de leur valeur totale et de leurs CV calculés sur une période de douze mois. De tels modèles peuvent être utilisés lorsqu'on doit estimer les CV de nouvelles caractéristiques ou pour fournir des estimations plus précises de la fiabilité des estimations fondées sur des données antérieures. Le problème de l'évaluation de la validité de l'ajustement des modèles est également examiné dans ce document.

## 1. INTRODUCTION

Ce document présente les résultats d'une étude d'évaluation des modèles de calcul des coefficients de variation (CV) d'estimations de caractéristiques de l'enquête sur la population active du Canada (EPA). L'EPA est une enquête mensuelle menée auprès des ménages, qui est basée sur un plan d'échantillonnage aréolaire stratifié à plusieurs degrés et dont l'échantillon comprend environ 55 000 ménages.

On estime chaque mois les CV d'un ensemble de caractéristiques en appliquant la méthode d'estimation de la variance de Keyfitz selon une approximation de la série de Taylor [4], [5].

<sup>1</sup> Exposé présenté à l'assemblée annuelle de l'American Statistical Association, en août 1981, à Détroit.

<sup>2</sup> P.D. Ghangurde, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.

En conclusion, j'aimerais aborder ce qui doit, selon moi, devenir prioritaire, soit l'analyse statistique de nos bases de données en vue d'en extraire toute l'information et d'en définir le sens et la portée pour en tirer vraiment des applications pratiques.

Depuis trop longtemps à Statistique Canada, nous nous sommes bornés à publier des myriades de chiffres sans en tirer d'indicateurs. Nous avons donc laissé aux autres le soin d'extraire l'or du minerai. Bien que nous comptions des réussites à ce chapitre, cette lacune a coûté cher à Statistique Canada et à tous ceux qui oeuvrent dans le domaine de la santé. En effet, nous avons privé les utilisateurs d'une ressource extrêmement précieuse, à savoir les moyens de leur permettre d'améliorer les conditions de vie.

J'ai du mal à prévoir l'incidence que ces changements auront sur l'esprit de collaboration qui doit animer les nombreux organismes qui participent à l'élaboration et à la mise à jour d'un ensemble exhaustif et nécessairement complexe de statistiques sur la santé. Permettez-moi seulement de dire qu'au Canada, en dépit de toutes les pressions exercées, cet esprit ne s'est pas érodé, bien au contraire. J'envisage donc l'avenir avec beaucoup d'optimisme, mais j'ajouterais que les organismes statistiques nationaux devront repenser leur rôle en fonction des besoins dont je viens de vous entretenir. Il me semble évident que notre vocation sera non plus de produire des statistiques, mais d'assurer la coordination des travaux statistiques de tous les organismes participants.

- Il est évident que lorsqu'ils seront en mesure de produire des statistiques à partir des données contenues dans leurs propres documents administratifs, ils seront sans doute moins disposés à répondre à nos questionnaires. Par conséquent, il nous faudra aller chercher les informations à même leurs documents. Il pourrait en résulter une diminution de la production statistique à cause d'un manque de comparabilité des données entre les organismes ou les différents paliers administratifs.
- Ces organismes acquerront donc beaucoup d'autonomie et seront moins portés à se tourner vers l'extérieur pour obtenir des statistiques.
- Par conséquent, ils se feront une idée beaucoup plus juste du milieu dans lequel ils évoluent.
- Ces mêmes organismes seront également d'importants utilisateurs de statistiques, en particulier en ce qui a trait à leur structure interne et à leur secteur de compétence.
- Nous pouvons cependant entrevoir l'apparition de nombreux producteurs de statistiques. Dans bien des cas, il s'agira d'organismes qui se considéreront comme des opérateurs de SIG. Il n'en demeure pas moins que la statistique occupera une place prépondérante dans leurs activités.

Je pense en particulier aux dossiers administratifs concernant les régimes d'assurance-maladie provinciaux. On y enregistre annuellement au delà de trente millions de cas où les médecins canadiens doivent donner des soins médicaux primaires. Nous avons déjà démontré les possibilités que ces bases de données offrent sur le plan statistique et nous formulons actuellement des propositions concernant leur exploitation au cours des prochaines années. L'affectation des crédits nécessaires peut cependant présenter certaines difficultés.

Les nouveaux besoins devraient déterminer l'orientation des travaux statistiques. Dans les faits, ce sont les progrès réalisés dans le domaine de l'information qui décident de cette orientation. L'ordinateur exerce une influence prépondérante sur la statistique de la santé et, grâce à lui, des changements profonds vont s'opérer au chapitre de la production et de l'utilisation des statistiques.

L'apparition de l'ordinateur a permis aux gestionnaires, aux spécialistes de la santé, aux administrateurs hospitaliers, aux gouvernements des États, aux administrations provinciales, aux universités et à diverses associations d'utiliser des données sous une forme statistique. Dans la fonction publique fédérale, l'ordinateur est venu modifier le traitement des données ainsi que leurs applications. En outre, les responsables de nombreux organismes oeuvrant dans le domaine de la santé utilisent maintenant les données informatisées, non seulement dans le cadre de recherches, mais aussi afin de mieux comprendre ce secteur d'activité et de prendre des décisions en toute connaissance de cause. Autrefois, les données disponibles étaient peu, voire pas du tout consultées.

J'exagérerais à peine en disant qu'il n'y a pas si longtemps, pratiquement seuls les organismes statistiques nationaux pouvaient s'offrir des ordinateurs de grande puissance. Les temps ont beaucoup changé, car aujourd'hui, les grands et petits organismes peuvent disposer sans trop de frais d'installations informatiques très complexes et faciles à exploiter. Les changements qu'entraîne une utilisation répandue de l'ordinateur seront profonds et je crois que nous n'en saisissons pas encore toute la portée.



proche avenir, étant donné que cette enquête permettra de connaître les effets de notre mode de vie sur notre santé et qu'elle aidera la population et les gouvernements à réagir.

Nous venons de voir qu'au Canada, il a été souvent difficile de satisfaire les besoins des utilisateurs par le biais d'enquêtes auprès des ménages. Toutefois, il existe un domaine où nous avons su prévoir ces besoins et avons réussi à y répondre efficacement. Je pense à nos études

certaines facteurs démographiques, sociaux, professionnels et écologiques pour la santé des citoyens. Grâce à la prévoyance et à la ténacité du personnel responsable de la statistique de l'état civil, qui travaille en collaboration avec quelques personnes-ressources de Statistique Canada et d'ailleurs, nous possédons maintenant un fichier informatique sur la mortalité où tous les décès survenus au Canada depuis trente ans sont

stockés selon cause du décès. De plus, nous disposons d'un système général de couplages des données qui sert à relier aux données du fichier sur la mortalité les groupes de population soumis aux facteurs énumérés plus haut. Il est également possible d'effectuer le couplage avec le fichier sur l'incidence du cancer qui, bien qu'encore incomplet, est tout de même assez important puisqu'il contient des données recueillies au cours d'une période de dix ans.

Mme Martha Smith, chef de la Section de la santé environnementale et professionnelle, se rendra en Ecosse avant la fin du mois pour y faire un exposé sur la question, dans lequel elle présentera une bibliographie composée surtout de documents canadiens. La transcription de cet exposé est offerte sur demande. J'aimerais souligner que Mme Smith et M. John Silans, chef de la Section de la statistique de l'état civil et des registres des maladies, sont dans l'assistance aujourd'hui.

Je me permets de signaler d'autres sources de données qui pourraient contribuer à l'orientation future de l'appareil statistique de la santé au Canada. En effet, certaines bases de données sont très volumineuses et, vrai semblablement, très précieuses, mais, en général, elles ne sont pas beaucoup utilisées pour l'établissement de statistiques nationales.



- les sources de renseignements existantes,

- les changements technologiques en matière de traitement et d'analyse

des données,

- la collaboration entre les organismes intéressés et enfin

- les budgets consentis.

Il va de soi que ces facteurs décident encore aujourd'hui et décideront

demain des travaux que nous pouvons mener à bonne fin.

Ces facteurs changent et se recourent de sorte qu'il se peut qu'au cours de périodes données, certains aient plus d'influence que d'autres sur le cours et la teneur de nos travaux.

Quoi qu'il en soit, au fil des ans, la statistique de la santé a beaucoup évolué au Canada et il semble évident que dans l'avenir, cette évolution ira en s'accélérant. Ce sont les besoins des utilisateurs qui doivent orienter les travaux statistiques et c'est effectivement ce qui se produit même si nous n'obtenons pas toujours les résultats escomptés. J'ai déjà souligné que les statistiques sur les décès au Canada étaient surtout compilées à partir des données sur la morbidité hospitalière. Toutes les personnes intéressées s'accordaient pour dire que cet état de choses n'était pas satisfaisant. Pour corriger cette situation, l'administration fédérale a établi, il y a quelques années, une enquête permanente sur la santé nationale qui a fait l'objet de travaux de planification détaillés et de tests rigoureux tant en ce qui a trait au cadre théorique qu'en ce qui concerne les méthodes

utilisées. Des données ont pu être recueillies pendant dix mois seulement, les compressions budgétaires imposées dans toute l'administration fédérale ayant entraîné l'abandon de l'enquête. Les premiers résultats viennent d'être publiés et la base de données constituée promet d'être une source de renseignements précieuse. Elle constituera en outre un instrument très utile aux responsables de la prise de décisions en matière de santé. Bien entendu, il faudra utiliser ces données avec beaucoup de prudence, étant donné

qu'elles portent sur une très courte période. A l'heure actuelle, nous ne sommes pas encore en mesure de prévoir quand l'Enquête Santé Canada sera rétablie. Toutefois, j'ai bon espoir que les travaux reprendront dans un

Santé Canada, ministères provinciaux de la santé, les établissements de santé, les associations hospitalières et enfin les services d'archives responsables des registres de l'état civil.

Je reviendrai plus loin à ces trois grandes questions que interviennent dans le domaine de la statistique de la santé, c'est-à-dire les renseignements qu'elle nous fournit et ceux qu'elle ne nous fournit pas, les points forts et les lacunes des sources de renseignements consultées, ainsi que le niveau de collaboration existant entre les organismes intéressés.

Comment expliquer les caractéristiques de notre statistique de la santé? Si l'on examine son évolution depuis une soixantaine d'années, j'estime qu'il peut être instructif d'y réfléchir. Il est impossible d'évaluer, en toute objectivité, de quelle façon a été établi l'ordre des priorités dans le passé. D'ailleurs, même aujourd'hui, le choix des priorités comporte une certaine part de subjectivité. Bref, dans le passé, je crois que, d'une part, nous avons cherché à répondre aux nouveaux besoins formulés par les utilisateurs ou exprimés dans les rapports de commissions royales d'enquête et que, d'autre part, nous avons pressenti ces nouveaux besoins et cherché à les satisfaire en exploitant les sources de renseignements existantes parce qu'elles offraient des possibilités intéressantes. Ces sources de renseignements représentaient pour nous le filon d'or que le prospecteur cherchait à découvrir par hasard. Souignons aussi que nous avons dû suivre le courant amorcé par les nombreux progrès technologiques et que nous avons en tirer profit. En outre, le climat de collaboration dans lequel nous travaillons a influencé aussi la nature de nos travaux. Enfin, les ressources à notre disposition, soit les ressources financières et humaines, et le matériel de traitement des données, nous ont permis d'entreprendre certains travaux mais pas d'autres.

En résumé, l'importance de nos réalisations a été fonction de quelques facteurs suivants:

- Les besoins qui étaient portés à notre attention et ceux que nous avions pressentis,

Pour produire des statistiques sur la santé, nous utilisons, aujourd'hui comme par le passé, deux grandes sources de renseignements. La première regroupe les établissements de santé, en particulier les hôpitaux généraux et psychiatriques. A partir des données contenues dans leurs dossiers, nous pouvons établir un éventail de statistiques sur les caractéristiques mêmes de ces établissements ainsi que sur leurs patients et les maladies soignées. La statistique hospitalière du Canada compte parmi les plus détaillées et les plus complètes au monde.

Les registres de l'état civil (registres des naissances, des mariages et des décès) constituent notre seconde source de renseignements. Nous en tirons d'importantes statistiques sur les causes de décès.

Des statistiques très variées sont produites à partir de ces deux sources de données précieuses, et certaines statistiques importantes sont tirées d'autres sources, notamment des registres des maladies à déclaration obligatoire et, dans le cas du nombre de cas de cancer, des registres provinciaux des tumeurs. J'ai apporté quelques exemplaires du Répertoire de données de la santé. Les personnes qui ne peuvent en obtenir un aujourd'hui pourront m'écrire à Statistique Canada pour que je leur en fasse parvenir un par la poste.

Les questions que je désire aborder cet avant-midi sont les suivantes :

- Premièrement, nos statistiques ne portent malheureusement que sur les personnes qui s'adressent à des établissements de santé pour s'y faire soigner.

- Deuxièmement, parce qu'elles sont tirées de résultats d'enquêtes ou de documents administratifs, ces statistiques, même si elles sont très utiles, comportent néanmoins des lacunes.

- Troisièmement, l'existence d'un volume aussi considérable de données témoigne bien de la collaboration étroite et constante qui existe depuis de nombreuses années entre Statistique Canada, l'en-étre et

LA STATISTIQUE DE LA SANTÉ AU CANADA:  
RÉTROSPECTIVE ET JALONS POUR L'AVENIR<sup>1</sup>

Lorne Rowebottom<sup>2</sup>

L'auteur décrit brièvement les facteurs qui déterminent la production de statistiques dans le domaine de la santé au Canada. Il insiste surtout sur les différentes sources de données et sur la longue tradition de collaboration entre les plusieurs organismes impliqués dans la production d'information dans le domaine de la santé.

Monsieur le président, je tiens tout d'abord à vous dire combien je suis heureux de faire partie de cette table ronde, car cela me permet de rendre hommage à Mme Dorothy Rice et à ses collègues du National Center for Health Statistics qui, depuis vingt-cinq ans déjà, mène des enquêtes dans le domaine de la santé. Les membres de Statistique Canada, en particulier ceux de la Division de la santé, ont toujours admiré le travail accompli par le personnel du Centre et ils se joignent à moi pour féliciter Mme Rice.

Notre président m'a demandé d'examiner des questions qui traduisent non seulement la situation actuelle dans le domaine de la santé au Canada, mais qui peuvent également être représentatives des préoccupations des statisticiens à l'étranger. J'espère que le tableau très général que j'ai choisi de brosser sur les tendances qui se manifestent dans le secteur de la santé au Canada et sur les facteurs qui peuvent en influencer l'évolution répond à ses vœux et saura, par conséquent, vous intéresser.

<sup>1</sup> Allocution prononcée à l'assemblée annuelle de l'American Statistical Association tenue à Détroit, en août 1981

<sup>2</sup> Lorne Rowebottom, Statisticien en chef adjoint, Direction de la statistique des institutions et de l'agriculture, Statistique Canada.



# BIBLIOGRAPHIE

- [1] Frankel, M.R. (1971), Inference from Survey Samples, University of Michigan, Ann Arbor.
- [2] Freeman, D.H. Jr., and Koch, G.G. (1976), "An Asymptotic Covariance Structure for Testing Hypotheses on Raked Contingency Tables from Complex Sample Surveys", Proc. Amer. Statist. Ass. (Social Statistics Section), Part 1, 330-335.
- [3] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankya, Series C, 37, 117-132.
- [4] Hajek, J. (1960), "Limiting Distributions in Simple Random Sampling from a Finite Population", Publ. Math. Inst. Hung. Acad. Sci., 5, 361-374.
- [5] Imrey, P.B., Koch, G.G., Stokes, M.E. (1981-1982), "Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression; Part I: Historical and Methodological Overview. Part II: Data Analysis", International Statistical Review. A paratre.
- [6] Kish, L., and Frankel, M.R. (1974), "Inference from Complex Samples", J. Roy. Statistic. Soc. B, 36, 1-22.
- [7] Madow, W.G. (1948), "On the Limiting Distribution of Estimates Based on Samples from Finite Universes", Ann. Math. Statist., 19, 535-545.
- [8] Särndal, C.E. (1978), "Design-based and Model-based Inference in Survey Sampling", Scand. J. Statist. 5, 27-52.
- [9] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Amer. Statist. Assoc. (Social Statistics Section, 11-8.
- [10] Woodruff, R.S. (1977), "A Simple Method for Approximating the Variance of a Complicated Estimate", J. Amer. Statist. Assoc. 66, 41-414.



#### 4. DISCUSSION

Les méthodes décrites dans le présent article s'appliquent à certains modèles précis, à titre d'exemple, voir Fuller (1975) et Freeman et Koch (1976). Les résultats généraux n'ont toutefois pas été présentés de façon explicite. Il est possible d'utiliser de nombreux programmes statistiques et types pour estimer les paramètres des modèles décrits, mais les variances et les tests d'hypothèses compris dans ces programmes ne seront pas valides. Les résultats présentés dans cet article sont fondés sur l'hypothèse que les estimateurs sont asymptotiquement normaux. Il importe d'effectuer des études empiriques sur la validité de ces approximations.

D'autres méthodes d'estimation de bon nombre de paramètres décrits dans cet article sont présentées par Imrey, Koch et Stokes (1981, 1982). Leur méthode de régression asymptotique fonctionnelle s'inscrit également dans le cadre général de la présente analyse, quant à l'estimation et la dérivation des variances.

Cette expression peut quelquefois être simplifiée de la manière suivante. Si on peut supposer que  $\tilde{N}/\tilde{N}^T \hat{=} \tilde{p}(\tilde{B})$ , alors pour  $\tilde{\pi} = \tilde{N}/\tilde{N}^T \hat{1}$  nous avons :

$$\tilde{V}[\tilde{\pi}] \hat{=} (\tilde{N}^T \hat{1})^{-2} (\tilde{I} - \tilde{p}(\tilde{B}) \hat{1}^T) \tilde{V}[\tilde{N}] (\tilde{I} - \hat{1}^T \tilde{p}(\tilde{B})^T),$$

de sorte que

$$\tilde{V}[\tilde{B}] \hat{=} (\tilde{A}^T \tilde{H}(\tilde{B}) \tilde{A})^{-1} \tilde{A}^T \tilde{V}[\tilde{\pi}] \tilde{A} (\tilde{A}^T \tilde{H}(\tilde{B}) \tilde{A})^{-1}. \quad (3.11)$$

En outre, la matrice des covariances de  $\tilde{p}(\tilde{B})$ , soient les probabilités estimées par case, nous est donnée par :

$$\tilde{V}[\tilde{p}(\tilde{B})] = \tilde{H}(\tilde{B}) \tilde{A} \tilde{V}[\tilde{B}] \tilde{A}^T \tilde{H}(\tilde{B}).$$

Les estimateurs de  $\tilde{V}[\tilde{B}]$  et de  $\tilde{V}[\tilde{p}(\tilde{B})]$  sont des expressions semblables dans lesquelles  $\tilde{N}$  et  $\tilde{B}$  sont remplacés respectivement par  $\tilde{N}$  et  $\tilde{B}$ . On pourrait alors déterminer directement la valeur de  $\tilde{V}[\tilde{N}]$ . Dans le cas où q est relativement plus grand que r, il serait plus efficace de procéder de la façon suivante. Supposons que

$Y_{ki} = 1$  si la k-ième unité appartient à la i-ième catégorie  
 $= 0$  dans tout autre cas,

par  $k=1, \dots, N; i=1, \dots, q$ . Soient  $\tilde{Y}^T = (Y_{k1}, \dots, Y_{kq})$ , et  $\tilde{W}_k = \tilde{A}^T [\tilde{I} - \tilde{p}(\tilde{B}) \hat{1}^T] \tilde{Y}_k$ .

Alors

$$\tilde{V}[\tilde{B}] \hat{=} (\tilde{N}^T \hat{1})^{-2} (\tilde{A}^T \tilde{H}(\tilde{B}) \tilde{A})^{-1} \tilde{V}(\tilde{W}) (\tilde{A}^T \tilde{H}(\tilde{B}) \tilde{A})^{-1}.$$

Souignons que les méthodes décrites dans la présente section peuvent facilement être appliquées aux modèles de distribution multinomiale du produit, dans lesquels il existe un modèle linéaire logarithmique pour  $\{N_{ij}\}$ , mais où les marges  $\{Z \cdot N_{ij}\}$  sont connues :

Supposons que  $\tilde{p}(\tilde{\beta})^T = [p_1(\tilde{\beta}), \dots, p_q(\tilde{\beta})]$  et que  $\tilde{N}^T = (N_1, \dots, N_q)$ , où  $N_i$  représente le nombre d'individus dans la  $i$ -ième catégorie. Si la population est générée à partir d'une distribution multinomiale de probabilité  $\tilde{p}(\tilde{\beta})$ , l'estimateur à maximum du vraisemblance de  $\tilde{\beta}$ , donné par  $\tilde{\beta}$ , satisfait

$$\tilde{U} = \tilde{A}^T \tilde{N} - [\tilde{A}^T \tilde{p}(\tilde{\beta})] \tilde{1}^T \tilde{N} = 0,$$

où  $\tilde{A}$  est une matrice  $q \times r$  dont la  $i$ -ième ligne correspond à  $\tilde{a}_i^T$ . Nous considérons  $\tilde{\beta}$  comme notre paramètre d'intérêt de toute population finie donnée.

Soit  $\tilde{N}$  un estimateur convergent asymptotiquement normal de  $\tilde{N}$ ,  $V[\tilde{N}]$  la matrice des variances-covariances et  $V[\tilde{N}]$  la matrice estimée. Notre estimateur,  $\tilde{\beta}$ , satisfait:

$$\tilde{A}^T \tilde{N} - [\tilde{A}^T \tilde{p}(\tilde{\beta})] \tilde{1}^T \tilde{N} = 0. \quad (3.9)$$

Cette méthode d'estimation a été proposée par Freeman et Koch (1976). Elle est peut-être moins efficace que la méthode de régression asymptotique fonctionnelle de Imrey, Koch et Stokes (1981, 1982), mais il n'est pas nécessaire de calculer toutes les composantes de  $V[\tilde{N}]$  pour appliquer la formule (3.9).

Nous supposons que  $\tilde{D}(\tilde{\beta})$  correspond à  $\text{diag}[\tilde{p}(\tilde{\beta})]$  et  $\tilde{H}(\tilde{\beta}) = \tilde{D}(\tilde{\beta}) - \tilde{p}(\tilde{\beta}) \tilde{p}(\tilde{\beta})^T$ . Alors

$$\tilde{U} = \frac{\partial \tilde{U}}{\partial \tilde{\beta}} = -(\tilde{1}^T \tilde{N}) \tilde{A}^T \tilde{H}(\tilde{\beta}) \tilde{A}.$$

Par conséquent, la matrice de variance asymptotique pour  $\tilde{\beta}$  est donnée par:

$$V[\tilde{\beta}] = (\tilde{N}^T \tilde{1})^{-2} (\tilde{A}^T \tilde{H}(\tilde{\beta}) \tilde{A})^{-1} \\ \tilde{A}^T (\tilde{I} - \tilde{p}(\tilde{\beta}) \tilde{1}^T) V[\tilde{N}] (\tilde{I} - \tilde{1} \tilde{p}(\tilde{\beta})^T) \tilde{A} (\tilde{A}^T \tilde{H}(\tilde{\beta}) \tilde{A})^{-1}. \quad (3.10)$$

Nous pouvons également estimer la variance de  $\hat{R}^2$ . Si  $\tilde{W}_k^T(B, R^2) = [Y_k, Z_{k1} e_k, \dots, Z_{kp} e_k, Y_k (\sum_j Z_{kj} B_j - R^2 Y_k)]$  et  $\tilde{c}^T = [-2\hat{Y}(1-\hat{R}^2)/N, \hat{B}^T, 1]/(S_{YY}^{-1} \hat{Y}^2)$ , nous obtenons :

$$\hat{V}[\hat{R}^2] = \tilde{c}^T \tilde{V}[\tilde{W}(\hat{B}, R^2)] \tilde{c}. \quad (3.5)$$

Dans le cas où  $N$  est inconnu, par exemple lorsque les unités primaires d'échantillonnage sont des régions géographiques, nous obtenons l'équation supplémentaire suivante :

$$U_4 = N - \Sigma 1, \quad (3.6)$$

Après avoir ajouté les unités de la ligne et de la colonne appropriées à  $\tilde{J}$  et après inversion, nous pouvons estimer  $V[\hat{R}^2]$  de la façon suivante

$$\text{Nous posons : } \tilde{W}_k^T(\hat{B}, R^2) = [Y_k, Z_{k1} e_k, \dots, Z_{kp} e_k, Y_k (\sum_j Z_{kj} B_j - R^2 Y_k), 1]$$

$$\text{et que } \tilde{c}^T = [-2\hat{Y}(1-\hat{R}^2)/N, \hat{B}^T, 1, \hat{Y}^2(1-\hat{R}^2)/N^2]/(S_{YY}^{-1} \hat{Y}^2).$$

Nous pouvons alors déterminer  $\hat{V}[\hat{R}^2]$  à l'aide de l'équation (3.5) pour ces nouvelles valeurs de  $\tilde{W}_k(\hat{B}, R^2)$  et  $\tilde{c}$ .

### 3.4 Régression Logistique

Comme dans la section précédente, nous posons comme hypothèse que la matrice de données  $\tilde{X}$  peut être répartie en  $[\tilde{Z}|\tilde{Y}]$ , mais ici  $\tilde{Y}$  est un vecteur de 0 et des 1. Dans le cadre de l'analyse statistique classique, le modèle de régression logistique de  $\tilde{Y}$  conditionnel à  $\tilde{Z}$  implique que les valeurs de  $Y_1, \dots, Y_N$  sont indépendantes et que  $Pr(Y_k = 1) = p_k(\tilde{g})$ , où

$$p_k(\tilde{g}) = \frac{\exp(\tilde{g}^T \tilde{Z}_k)}{\exp(\tilde{g}^T \tilde{Z}_k) + 1} \quad (3.7)$$

En supposant que  $\tilde{B}$  est l'estimateur du maximum de vraisemblance de  $\tilde{g}$ , alors  $\tilde{B}$  satisfait

$$\tilde{U} = \tilde{Z}^T \tilde{P}(\tilde{B}) - \tilde{Z}^T \tilde{Y} = 0, \quad (3.8)$$

où  $\tilde{P}(\tilde{B})^T = [p_1(\tilde{B}), \dots, p_N(\tilde{B})]$ .

Pour une population finie donnée, nous définissons  $\tilde{B}$  comme notre paramètre d'intérêt.

Nous choisissons  $\tilde{C}(\tilde{B})$  comme notre estimateur de  $\tilde{Z}^T \tilde{P}(\tilde{B})$  et  $\tilde{S} \tilde{Z} \tilde{Y}$  comme notre estimateur de  $\tilde{Z}^T \tilde{Y}$ . Par conséquent,  $\tilde{B}$  satisfait  $\tilde{C}(\tilde{B}) = \tilde{S} \tilde{Z} \tilde{Y}$ . Ces

équations doivent habituellement être résolues de façon itérative. Posons également

$$\tilde{J} = \frac{\partial \tilde{U}}{\partial \tilde{B}}.$$

La  $(i, j)$ -ième composante de  $\tilde{J}$  est  $\sum_k Z_{ki} Z_{kj} p_k(\tilde{B}) [1 - p_k(\tilde{B})]$ . Nous notons l'estimateur de  $\tilde{J}$  par  $\tilde{J}$ .

Pour estimer la variance de  $\tilde{B}$ , nous posons :

$$\tilde{M}_k^T = (Z_{k1} e_k, \dots, Z_{kN} e_k)$$

où  $e_k = p_k(\tilde{B}) - Y_k$ . L'estimateur de  $\tilde{V}[\tilde{B}]$  est donné par :

$$\tilde{J}^{-1} \tilde{V}(\tilde{M}) \tilde{J}^{-1}.$$

### 3.5 Modèles linéaires logarithmiques pour des données qualitatives

Supposons que chaque membre de la population appartienne à une seule catégorie d'un nombre  $q$  de catégories distinctes. Un vecteur  $a_i$  de  $r \times 1$  est associé à la catégorie  $i$  de sorte que la proportion d'individus faisant partie de la  $i$ -ième catégorie est approximativement :

$$p_i(\tilde{g}) = \frac{\exp(a_i^T \tilde{g})}{\sum_j \exp(a_j^T \tilde{g})}.$$



$$\hat{R}_2 = 1 - \frac{\hat{S}_{yy} - \hat{B}^T \hat{S}_{zy} \hat{Z}^{-1} \hat{Y}_2}{\hat{S}_{yy} - N^{-1} \hat{Y}_2^2} \quad (3.3c)$$

et

$$\tilde{J} = a(\tilde{Z}, \tilde{Y}, \tilde{B}, \hat{R}_2, \theta) / a(\tilde{B}, \hat{R}, \theta) = \begin{bmatrix} \tilde{0}^T & \tilde{Z}^T \tilde{Z} & \tilde{Z}^T \tilde{Z} & -\tilde{Y}^T \tilde{Z} \\ 0 & \tilde{0} & \tilde{0} & SSY \\ 1 & \tilde{0} & 2\tilde{Y}^T(1-R^2) & \end{bmatrix},$$

$$\text{ou } \tilde{Y} = \theta / N.$$

Il résulte donc que

$$\tilde{J}^{-1} = \begin{bmatrix} \tilde{0} & (\tilde{Z}^T \tilde{Z})^{-1} & \tilde{0} \\ -2\tilde{Y}(1-R^2)/SSY & B^T/SSY & 1/SSY \\ 1 & \tilde{0}^T & 0 \end{bmatrix}.$$

Si nous supposons maintenant que  $\tilde{W}_k(B) = (Z_{k1} e_k, \dots, Z_{kp} e_k)$ , où  $e_k = Y_k - \sum_j Z_{kj} B_j$ , nous obtenons :

$$\tilde{V}[\tilde{B}] = (\tilde{Z}^T \tilde{Z})^{-1} \tilde{V}[\tilde{W}(\tilde{B})](\tilde{Z}^T \tilde{Z})^{-1}.$$

(3.4)

Il s'agit là d'une conséquence directe de l'équation (2.10). Notons que l'ensemble des vecteurs de  $\tilde{W}_k(\tilde{B})$  correspond à  $\tilde{U}_2$  dans l'équation (3.2b). Fuller (1975) obtient le même résultat avec un échantillon stratifié ou un échantillon stratifié à deux degrés.

Pour estimer (3.4), nous utilisons :

$$\tilde{V}[\tilde{B}] = \tilde{S}_{ZZ}^{-1} \tilde{V}[\tilde{W}(\tilde{B})] \tilde{S}_{ZZ}^{-1}.$$

Si  $\hat{U}(\hat{x}, R) = 0$ , nous obtenons

$$\hat{R} = \hat{x}_2 / \hat{x}_1. \quad (3.1)$$

Nous posons que  $W_k = X_{k2} - R X_{k1}$ .

Comme  $\hat{J}(\hat{x}, R) = -\sum X_{k1}^2$ , nous pouvons affirmer que  $V(\hat{R})$  peut s'approximer par  $V(\hat{W}) / (\sum X_{k1}^2)$ , estimé par  $V(\hat{W}) / \sum \hat{X}_1^2$ . Dans le cas d'un échantillonnage stratifié, le résultat obtenu serait le même que dans l'étude de Woodruff (1971).

### 3.3 Coefficients de régression et R

Supposons que notre matrice de données  $\tilde{X}$  est répartie en  $[\tilde{Z} | \tilde{Y}]$ , la première colonne de  $\tilde{Z}$  étant le vecteur des 1,  $\tilde{Y}$  étant un vecteur par  $N \times 1$ . Les paramètres d'intérêt  $\theta$ ,  $\tilde{B}$ , et  $R^2$  sont définis par:

$$U_1 = \theta - \tilde{Y}^T \tilde{1} = 0, \quad (3.2a)$$

$$\tilde{U}_2 = \tilde{Z}^T \tilde{Z} \tilde{B} - \tilde{Z}^T \tilde{Y} = 0, \quad (3.2b)$$

$$U_3 = (\tilde{Y}^T \tilde{Y} - N^{-1} \theta^2)(R^2 - 1) + \tilde{Y}^T \tilde{Y} - \tilde{Y}^T \tilde{Z} \tilde{B} = 0. \quad (3.2c)$$

Dans cet exemple,  $\tilde{B}$  représente le vecteur des coefficients de régression,

$R^2$  le coefficient de détermination multiple et  $\theta$ , le total de tous les  $Y$ . Prenons d'abord le cas où  $N$  est connu. Nous supposons que  $SSY = \tilde{Y}^T \tilde{Y} - N^{-1} \theta^2$ . Définissons également,  $\tilde{S}^{ZZ}$ , l'estimateur de  $\tilde{Z}^T \tilde{Z}$ ,  $S_{YY}$ , l'estimateur de  $\tilde{Y}^T \tilde{Y}$  et  $\tilde{S}^{ZY}$ , l'estimateur de  $\tilde{Z}^T \tilde{Y}$ . Par conséquent, nous avons:

$$\hat{\theta} = \bar{Y}, \quad (3.3a)$$

$$\hat{\tilde{B}} = \tilde{S}^{ZZ-1} \tilde{S}^{ZY}, \quad (3.3b)$$

### 3. EXEMPLES

#### 3.1 Introduction

Dans la présente partie, nous analysons en détail les applications des énoncés généraux formulés dans la deuxième partie au sujet de l'estimation des variances de certains estimateurs des paramètres de la population. Nous examinerons tout particulièrement les rapports, les coefficients de régression et les modèles linéaires logarithmiques se rapportant à des données qualitatives. D'autres modèles comme les modèles de probit pourraient être analysés de façon analogue.

Nous nous fondons généralement sur les notations suivantes. Si  $\tilde{W}_1, \dots, \tilde{W}_n$  représentent des valeurs de la population et que  $\tilde{W} = \sum \tilde{W}_k$ , lorsque nous tirons un échantillon  $\tilde{w}_1, \dots, \tilde{w}_n$ , nous avons donc un estimateur sans biais de  $\tilde{W}$  indiqué par  $\tilde{\hat{W}}$ . Nous supposons que  $\tilde{V}(\tilde{\hat{W}})$  représente la matrice des covariances de  $\tilde{\hat{W}}$  et que  $\tilde{V}(\tilde{\hat{W}})$  est un estimateur convergent de  $\tilde{V}(\tilde{W})$ . La forme particulière que prendra cet estimateur dépend du plan de d'échantillonnage, par exemple, stratifié à plusieurs degrés, ppt. avec remise, etc..

#### 3.2 Rapports

Supposons que nous nous intéressions au rapport  $R = \sum X_{k2} / \sum X_{k1}$ . Nous définissons

$$U(\tilde{X}, R) = \sum X_{k2} - R \sum X_{k1}.$$

Par conséquent, dans le cas d'un échantillonnage sans remise, nous observons que

$$U(\tilde{x}, R) = \sum \tilde{x}_{k2} - R \sum \tilde{x}_{k1}.$$

Supposons en outre que

$$\lim_{t \rightarrow \infty} [\text{rang} \{ \tilde{J}(\tilde{x}, \tilde{\theta}) \}] = \text{plim} [\text{rang} \{ \tilde{J}(\tilde{x}, \tilde{\theta}) \}] = p.$$

Nous définissons  $\hat{\theta}(t)$  pour satisfaire

$$\tilde{J}(\tilde{x}(t), \hat{\theta}(t)) = 0.$$

Grâce à un développement de la série de Taylor, nous obtenons

$$\tilde{J}(\tilde{x}(t), \hat{\theta}(t)) \approx -\tilde{J}(\tilde{x}(t), \tilde{\theta}(t)) + \tilde{J}(\tilde{x}(t), \tilde{\theta}(t)) - \tilde{J}(\tilde{x}(t), \tilde{\theta}(t)). \quad (2.9)$$

Comme le premier membre de l'équation (2.9) est asymptotiquement normal, il s'ensuit que

$$(\tilde{J}(\tilde{x}(t), \hat{\theta}(t)) - \tilde{J}(\tilde{x}(t), \tilde{\theta}(t)))$$

Soit asymptotiquement une loi  $N[0, \tilde{G}(\tilde{\theta}_0)]$ , ou

$$\tilde{S}(\tilde{\theta}) = \tilde{K}(\tilde{\theta}_0) \tilde{G}(\tilde{\theta}_0) [\tilde{K}(\tilde{\theta}_0)]^T.$$

Par conséquent,

$$\tilde{G}(\tilde{\theta}_0) = [\tilde{K}(\tilde{\theta}_0)]^T \tilde{S}(\tilde{\theta}_0) [\tilde{K}(\tilde{\theta}_0)]^T \quad (2.10)$$

et

$$n(t) [\tilde{J}(\tilde{x}, \tilde{\theta})]^T \tilde{S}(\tilde{\theta}) [\tilde{J}(\tilde{x}, \tilde{\theta})]^{-1} [\tilde{J}(\tilde{x}, \tilde{\theta})]^T. \quad (2.11)$$

est un estimateur convergent de  $\tilde{G}(\tilde{\theta}_0)$

En conclusion, lorsque la forme fonctionnelle de  $\tilde{U}(\tilde{x}, \tilde{\theta})$  et de  $\tilde{J}(\tilde{x}, \tilde{\theta})$  est spécifiée, il ne nous reste plus qu'à dériver la matrice  $\tilde{J}(\tilde{x}, \tilde{\theta}_0)$  et son estimateur  $\hat{J}(\tilde{x}, \tilde{\theta})$  pour pouvoir utiliser ces résultats.

$\hat{J}(\tilde{x}, \tilde{\theta})$  est un estimateur convergent de  $J(\tilde{x}, \tilde{\theta})$ .

Notre estimateur de  $\hat{\theta}$  est donné par  $\hat{\theta}$ , la solution à :

$$\hat{U}_1(\tilde{x}, \tilde{\theta}) = 0, \text{ puisque } i=1, \dots, p. \quad (2.8)$$

Nous supposons que la taille de l'échantillon est suffisamment grande pour que la solution de l'équation (2.8) soit unique dans  $\hat{\theta}$ . Nous verrons dans la prochaine section que la matrice des covariances de  $\hat{\theta}$  peut être estimée de façon convergente par :

$$[\hat{J}^{-1}(\tilde{x}, \tilde{\theta})] \hat{\Sigma}(\tilde{x}, \tilde{\theta}) [\hat{J}^{-1}(\tilde{x}, \tilde{\theta})]^T.$$

## 2.2 Théorie asymptotique

Conformément aux raisonnements asymptotiques de Madow (1948) et Hajek (1960), nous étudions une séquence de populations avec l'indice  $t$ , de tailles  $N(t)$  et à partir de données  $\tilde{X}(t)$ . Nous supposons que  $N(t) \rightarrow \infty$  quand  $t \rightarrow \infty$ . Pour la population  $t$ , nous tirons un échantillon de taille  $n(t)$  et faisons porter nos observations sur les données  $\tilde{x}(t)$ . Nous posons que  $v(t) = E(n(t))$  et nous supposons que  $\lim_{t \rightarrow \infty} v(t) = \infty$  et  $\lim_{t \rightarrow \infty} N(t) - v(t) = \infty$ . Pour tout  $\tilde{\theta}$  dans le voisinage de  $\tilde{\theta}_0(t)$ , nous supposons que

$$[v(t)]^{-1} [\hat{J}(\tilde{x}, \tilde{\theta}) - \hat{J}(\tilde{x}, \tilde{\theta}_0(t))] = o_p(1)$$

suit asymptotiquement une loi  $N[0, \tilde{S}(\tilde{\theta})]$ , où

$$\tilde{S}(\tilde{\theta}) = \lim_{t \rightarrow \infty} [v(t)]^{-1} \tilde{\Sigma}(\tilde{x}, \tilde{\theta}) / N(t)$$

existe. Nous supposons également l'existence de

$$K(\tilde{\theta}) = \lim_{t \rightarrow \infty} J(\tilde{x}, \tilde{\theta}) / N(t) \text{ et aussi}$$

$$\lim_{t \rightarrow \infty} \hat{J}(\tilde{x}, \tilde{\theta}) / N(t) = K(\tilde{\theta}).$$



Soit  $\tilde{\Sigma}(\tilde{X}, \tilde{\theta})$  la matrice pxp avec les éléments  $\sigma_{ij}(\tilde{X}, \tilde{\theta})$ , et  $\tilde{\xi}(\tilde{X}, \tilde{\theta})$  un estimateur convergent de  $\tilde{\Sigma}$ . Maintenant, pour tout  $\tilde{\theta}$  donné,

$$u_i(\tilde{X}, \tilde{\theta}) + v_i(\tilde{\theta}) = \sum_{k=1}^N u_i(\tilde{X}_k, \tilde{\theta}),$$

de sorte que les estimateurs  $\hat{u}_i(\tilde{X}, \tilde{\theta})$  et  $\hat{v}_i(\tilde{X}, \tilde{\theta})$  peuvent être spécifiés pour tout plan dans lequel nous pouvons dériver des estimateurs convergents asymptotiquement normaux des valeurs totales de la population, de même que des estimateurs convergents des variances des estimateurs des valeurs totales.

L'estimateur de Horvitz-Thompson pour (2.3) est

$$(2.4) \quad \sum_{k=1}^n \sum_{\tilde{\theta}=1}^n u_i(\tilde{X}_k, \tilde{\theta}) - \pi_k \pi_{\tilde{\theta}} / \pi_k \pi_{\tilde{\theta}} \pi_k.$$

Dans le cas d'un échantillon de taille déterminée, l'estimateur de Yates-Grundy pour (2.3) est:

$$(2.5) \quad \sum_{k > \tilde{\theta}} \left[ \frac{u_i(\tilde{X}_k, \tilde{\theta})}{u_i(\tilde{X}_{\tilde{\theta}}, \tilde{\theta})} - \frac{\pi_k}{\pi_{\tilde{\theta}}} \right] \left[ \frac{u_j(\tilde{X}_k, \tilde{\theta})}{u_j(\tilde{X}_{\tilde{\theta}}, \tilde{\theta})} - \frac{\pi_k}{\pi_{\tilde{\theta}}} \right] \cdot (\pi_k \pi_{\tilde{\theta}} - \pi_k \pi_{\tilde{\theta}}).$$

Soit  $\tilde{U}(\tilde{X}, \tilde{\theta})$  et  $\tilde{U}_i(\tilde{X}, \tilde{\theta})$  les vecteurs p-dimensionnels dont les composantes sont respectivement  $U_i(\tilde{X}, \tilde{\theta})$  et  $U_i(\tilde{X}, \tilde{\theta})$ , nous établissons que

$$(2.6) \quad \tilde{C}(\tilde{X}, \tilde{\theta}) = a \tilde{U}(\tilde{X}, \tilde{\theta}) / a \tilde{\theta}$$

$$(2.7) \quad \tilde{C}(\tilde{X}, \tilde{\theta}) = U(\tilde{X}, \tilde{\theta}) / a \tilde{\theta}$$

où  $\tilde{C}$  et  $\tilde{U}$  sont des matrices de dérivées partielles pxp. Supposons que les matrices sont des fonctions continues de  $\tilde{\theta}$  et que les dérivées partielles relatives à  $\tilde{\theta}_0$  existent dans le voisinage de  $\tilde{\theta}_0$ . Supposons également que

$$U_i(\tilde{x}, \tilde{\theta}_0) = \sum_{k=1}^K u_i(\tilde{x}_k, \tilde{\theta}_0) - v_i(\tilde{\theta}_0) = 0, \quad (2.1)$$

Nous supposons que les équations (2.1) définissent  $\theta_0$  uniquement dans  $\theta$ . Nous supposons également que  $u_i(\tilde{x}, \tilde{\theta})/a\tilde{\theta}$  et  $v_i(\tilde{\theta})/a\tilde{\theta}$  existent dans le voisinage de  $\tilde{\theta}_0$ . Un exemple simple de (2.1) est lorsque  $\theta_0$  représente un total de la population et que nous avons  $U(\tilde{x}, \theta_0) = \sum_{k=1}^K x_k - \theta_0$ . Dans ce cas,  $u(x_k, \theta_0) = x_k$  et  $v(\theta_0) = \theta_0$ .

Nous tirons un échantillon d'unités selon une distribution de probabilité définie en sur l'ensemble de tous les sous-ensembles non vides de  $\{1, \dots, N\}$ . Nous désignons par  $\tilde{x}_1, \dots, \tilde{x}_n$  les valeurs choisies de  $\tilde{x}_1, \dots, \tilde{x}_N$ . Nous supposons que pour tout  $\tilde{\theta} \in \theta$ , nous pouvons construire un estimateur convergent asymptotiquement normal de  $U_i(\tilde{x}, \tilde{\theta})$ . Nous désignons cet estimateur par  $\hat{U}_i(\tilde{x}, \tilde{\theta})$ . Par exemple, pour nombre de plans d'échantillonnage sans remise,

$$\hat{U}_i(\tilde{x}, \tilde{\theta}) = \frac{1}{n} \sum_{k=1}^n u_i(\tilde{x}_k, \tilde{\theta}) / \pi_k - v_i(\tilde{\theta}) \quad (2.2)$$

est un estimateur convergent asymptotiquement normal, où  $\pi_k$  représente la probabilité d'inclusion de la  $k$ -ième unité. Posons  $\sigma_{ij}(\tilde{x}, \tilde{\theta}) = \text{Cov}[\hat{U}_i(\tilde{x}, \tilde{\theta}), \hat{U}_j(\tilde{x}, \tilde{\theta})]$ . Par exemple, pour l'estimateur (2.2), nous constatons que

$$\sigma_{ij}(\tilde{x}, \tilde{\theta}) = \frac{1}{N} \sum_{k=1}^K u_i(\tilde{x}_k, \tilde{\theta}) u_j(\tilde{x}_k, \tilde{\theta}) (\pi_k - \pi_k^2) / \pi_k \pi_j, \quad (2.3)$$

où  $\pi_k$  représente la probabilité que la  $k$ -ième et la  $j$ -ième unité soient dans l'échantillon.

(1.1)

$$\tilde{X}'\tilde{X}\tilde{B} = \tilde{X}'\tilde{Y}$$

Cette conception des paramètres descriptifs est la même que celle qui a été considérée par Frankel (1971) et Kish et Frankel (1974)

Pour l'estimation de tels paramètres, on tient compte habituellement des coefficients de pondération de l'échantillonnage. Si nous indiquons par  $\Pi_1$  la probabilité que la  $i$ -ième unité de l'échantillon soit prélevée et si  $\tilde{\Pi} = \text{diag}(\pi_1, \dots, \pi_n)$ , alors l'estimateur pondéré du paramètre  $\tilde{B}$  satisfait:

$$\tilde{X}'\tilde{\Pi}^{-1}\tilde{X}\tilde{B} = \tilde{X}'\tilde{\Pi}^{-1}\tilde{Y} \quad (1.2)$$

où  $\tilde{X}$  et  $\tilde{Y}$  représentent respectivement des matrices  $n \times p$  et  $n \times 1$  dont les lignes correspondent aux lignes échantillonnées de  $\tilde{X}$  et de  $\tilde{Y}$ .

Supposons maintenant qu'un estimateur d'un paramètre d'une population peut être défini de la façon suivante:

$$\hat{\theta} = g(z_1, \dots, z_k), \quad (1.3)$$

où  $E(z_1) = z_1$ . Dans cette équation,  $\hat{\theta}$  est un estimateur de  $g(z_1, \dots, z_k)$ . Selon Tepping (1968) et Woodruff (1971), un développement de la série de Taylor pour  $\hat{\theta}$  donne:

$$V[\hat{\theta}] \approx V\left[\sum_{k=1}^k \left(\frac{\partial \hat{\theta}}{\partial z_k}\right) (z_k - z_k^0)\right] \quad (1.4)$$

Des exemples de ces formules appliquées à l'estimation des coefficients de régression (1.1) ont été fournis par Tepping (1968). Toutefois, pour le calcul des variances des coefficients de régression, les expressions qui découlent de l'équation (1.4) sont un peu plus complexes que celles qui ont été dérivées par Fuller (1975).

Les paramètres examinés dans cet article ne sont pas définis par une équation explicite comme l'équation (1.3), mais ils sont plutôt définis implicitement sous la forme  $\tilde{U}(\tilde{Z}, \tilde{\theta}) = 0$ . Un exemple simple d'une telle distinction pourrait être le paramètre du rapport :

$$R = \frac{\sum Y_k}{\sum X_k},$$

qu'on pourrait aussi définir implicitement sous la forme

$$\sum Y_k - R \sum X_k = 0.$$

Pour certains types de modèles, comme les modèles linéaires logarithmiques indirects ou les modèles de régression logistique, les paramètres peuvent être définis seulement en fonction de relation implicites. Bien qu'elle n'apparaîse encore sous une forme générale dans la littérature l'extension des résultats de Tepping (1968) est relativement simple à appliquer dans ce cas. Il existe toutefois quelques exemples précis de son application, voir notamment Fuller (1975) et Freeman et Koch (1976).

Le cadre général et les principaux résultats de notre étude sont exposés dans la deuxième partie du présent article et des exemples d'un certain nombre de modèles sont présentés dans la troisième partie.

## 2. CADRE GÉNÉRAL

### 2.1 Cadre

Les unités de la population sont désignées  $1, \dots, N$ . Un vecteur de données  $q$ -dimensionnel  $\tilde{X}_i$  est associé à la  $i$ -ième unité. Nous avons un espace des paramètres  $\theta \in R^p$ . Le paramètre  $\tilde{\theta}_0 = (\theta_{10}, \dots, \theta_{p0})$  est défini par les équations :

sous-population. Pour cette raison et pour d'autres considérations d'ordre pratique, le plan d'enquête comprend rarement un échantillon aléatoire simple, mais plutôt un échantillon stratifié, souvent à plusieurs degrés, pouvant présenter des probabilités inégales à certains degrés.

Aussi, on s'est beaucoup interrogé sur le bien-fondé de l'utilisation de coefficients de pondération de l'échantillonnage dans les inférences faites au sujet de ces paramètres de modèles (voir notamment Särndal, 1978). Pour pouvoir répondre à une telle question, il faut d'abord déterminer si un modèle de superpopulation convient à toutes les unités de la population. Dans l'affirmative, les inférences portant sur les paramètres de la superpopulation constituent souvent la première occupation. Il s'agit alors d'inférences basées sur des modèles, c'est-à-dire que, pour un échantillon donné, elles ne dépendent pas des coefficients de pondération de l'échantillonnage.

La question qui nous vient à l'esprit est la suivante, si le modèle de superpopulation ne convient pas, à quoi correspondent les paramètres que nous estimons? Il nous faut admettre que dans de nombreuses études, particulièrement dans le domaine des sciences sociales, le modèle utilisé (par exemple la régression linéaire) n'est qu'une approximation utile de l'univers réel et que les paramètres de ce modèle (par exemple les corrélations et les corrélations partielles) ont plus souvent pour objet de faciliter la compréhension des interdépendances approximatives des variables que de donner lieu à une interprétation scientifique particulière. Par conséquent, les paramètres que nous estimons ne se rapportent pas nécessairement à un modèle de superpopulation réel; ils ont plutôt un caractère descriptif.

Dans cet article, nous partons du principe que nous cherchons à faire des inférences au sujet de ces paramètres "descriptifs" de la population. À titre d'exemple, supposons que  $\tilde{X}$  et  $\tilde{Y}$  désignent respectivement des matrices  $N \times p$  et  $N \times 1$ , chaque ligne de  $\tilde{X}$  et  $\tilde{Y}$  correspondant à un individu différent de la population. Nous nous intéressons au paramètre descriptif,  $\tilde{B}$ , un vecteur  $p \times 1$  qui satisfait l'équation:



# LES VARIANCES D'ESTIMATEURS ASYMPTOTIQUEMENT NORMAUX BASES SUR DES ENQUÊTES COMPLEXES

David A. Binder<sup>1</sup>

Nous discutons du problème de la spécification et de l'estimation de la variance de paramètres estimés basés sur des plans d'échantillonnage complexes provenant de populations finies. Les résultats présentés dans cet article sont particulièrement utiles lorsque les estimateurs des paramètres ne sont pas définis explicitement comme étant une fonction des autres statistiques de l'échantillon. Nous montrons comment des résultats peuvent s'appliquer à la régression linéaire, à la régression logistique et aux modèles linéaires logarithmiques de tables de contingence.

## I. INTRODUCTION

Nous avons observé ces dernières années une tendance croissante à utiliser des données d'enquête pour estimer les paramètres de modèle classiques, tels que les paramètres de régression, les fonctions discriminantes et les paramètres de logit et de probit. Toutefois, dans nombre d'enquêtes, l'objet premier est l'estimation des valeurs moyennes, des valeurs totales, des tendances et autres caractéristiques d'une population ou d'une

<sup>1</sup> D.A. Binder, Division des méthodes d'enquête-institution et agriculture, Statistique Canada.

Références (suite)

- Tulder, J.J.M. van, 1977, Op de grens van non-response, jaarboek van de Nederlandse Vereniging van Marktonderzoekers 1977, pp 43-52
- Widdershoven, M. & J. van den Berg, 1980, Non-respons bij twee "persoons- en gezinsenquêtes", in: CBS Select 1 (Staatsuitgeverij, The Hague).

## Références

- Benzécri, J.P., 1976, L'analyse des Données (Dunod, Paris)
- Bethlehem, J.G. & H.M.P. Kersten, 1981, Graphical Methods in Non-response Analysis and Sample Estimation (Staatsuitgeverij, The Hague)
- Bishop, Y.M.M., S.E. Fienberg & P.W. Holland, 1975, Discrete Multivariate Analysis (MIT Press, Cambridge)
- Chapman, D.W., 1976, A survey of Non-response Imputation Procedures, Proceedings of the American Statistical Association, social statistics section, pp 245-251
- Hansen, M.H. & W.N. Hurwitz, 1946, The Problem of Non-response in Sample Surveys, Journal of the American Statistician 41, pp 517-529
- Holt, D. & T.M.F. Smith, 1979, Post Stratification, Journal of the Royal Statistical Society, series A, 142, pp 33-46
- Kass, G.V., 1980, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics 29, pp 119-127
- Kish, L., 1967, Survey Sampling (Wiley, New York)
- Morgan, J.N. & J.A. Sonquist, 1963, Problems in the Analysis of Survey Data, Journal of the American Statistical Association 58, pp 415-434
- Politz, A. & W. Simmons, 1949, An Attempt to get the Not-at-homes into the Sample without Callbacks, Journal of the American Statistical Association 44, pp 9-31
- Schmid, C.F., 1954, Handbook of Graphic Presentation (Ronald Press, New York)

## 5.6. La principale question

Si la méthode de Hansen et Hurwitz est trop coûteuse, on peut recourir en remplacement à la méthode de la principale question. Dans la plupart des enquêtes il y a souvent une question de base importante autour de laquelle l'enquête a été centrée. Si au cours des travaux sur le terrain on a des problèmes pour faire remplir le questionnaire, l'enquêteur peut essayer d'obtenir une réponse portant uniquement sur la principale question. On peut même essayer de recourir à cette méthode ultérieurement par lettre ou par téléphone. Cette technique sera testée sous peu dans une des enquêtes du BCS.

## 6. CONCLUSIONS

Par suite de l'augmentation des taux de non-réponse qui a été constatée ces dernières années, il importe d'effectuer des recherches approfondies en ce qui concerne l'incidence de la non-réponse sur la qualité de l'enquête.

Les ouvrages techniques exposent un assez grand nombre de méthodes d'ajustement qui visent toutes à réduire la distorsion imputable à la non-réponse. Une étude comparative de ces méthodes doit fournir des réponses décisives au sujet de leurs avantages.

Les grandes différences qui existent en ce qui concerne l'objectif, la conception et l'exécution des enquêtes empêchent d'interpréter comme il convient les différences constatées dans les chiffres de non-réponse. Il est par conséquent nécessaire de mettre en place un cadre théorique qui permette d'effectuer une comparaison appropriée.

Bien entendu, la suppression de la non-réponse au cours des travaux sur le terrain restera un sujet important.

### 5.3. Ajustement pour tenir compte des "absents du foyer"

La méthode bien connue de Politz et Simmons (1949) tente de remédier à la distorsion imputable aux absents du foyer en estimant la probabilité qu'il y a de trouver une personne chez elle. Pour cela on demande par exemple aux répondants combien de fois ils étaient chez eux à l'heure de l'entrevue au cours des jours précédents. La probabilité qu'ils soient chez eux, calculée de cette manière, peut construire un modèle expliquant la relation entre la variable étudiée et la probabilité de trouver les répondants chez eux. L'extrapolation de ce modèle au groupe des absents du foyer donne parfois davantage de renseignements au sujet de ce groupe.

### 5.4. Ajustement pour tenir compte des refus

Il est possible de déterminer dans quelle mesure les gens sont disposés à coopérer à l'enquête (voir Van Tulder, 1977). En utilisant ce renseignement, on peut appliquer une méthode d'ajustement analogue à celle qui est utilisée pour les absents du foyer. De plus, la volonté de coopérer est une mesure du climat dans lequel se déroule l'enquête. La construction d'une échelle de valeurs qui puisse renseigner à ce sujet est probablement plus difficile que dans le cas de l'ajustement appliqué aux absents du foyer.

### 5.5. Double échantillonnage

Pour recueillir davantage de renseignements au sujet des non-répondants, Hansen et Hurwitz (1946) ont proposé de choisir un échantillon parmi les non-répondants. Des enquêteurs spécialement formés essaient d'obtenir malgré tout les renseignements (ou une partie des renseignements) qui manquent. Les contraintes imposées par le manque de temps et d'argent empêchent souvent de recourir au double échantillonnage.



## 5.2. Imputation

L'imputation permet de résoudre le problème de la non-réponse en remplaçant les observations manquantes par des valeurs tirées des données provenant des réponses obtenues à l'occasion de l'enquête en cours, ou bien de données provenant d'une enquête antérieure. Si la structure des réponses, dans l'enquête en cours, est analogue à celle de l'enquête antérieure, les résultats des deux modes d'imputation seront à peu près les mêmes. L'imputation peut se faire de différentes manières. En voici quelques-unes :

- (1) imputation de la valeur observée pour un répondant pris au hasard
- (2) imputation de la valeur moyenne observée pour l'ensemble des répondants
- (3) imputation de la valeur observée pour un répondant pris au hasard dans le même sous-groupe
- (4) imputation de la valeur moyenne observée pour les répondants du même sous-groupe
- (5) imputation d'une valeur obtenue par ajustement d'un modèle
- (6) imputation de limites supérieures ou inférieures.

Les méthodes (1) et (2) ne réduisent pas la distorsion. Les méthodes (3) et (4) ressemblent à la pondération par sous-groupes. L'effet de la méthode (5) dépend pour beaucoup de la précision du modèle et de la qualité des hypothèses sur lesquelles il est fondé. La méthode (6) donne une idée de ce que pourrait être l'erreur si aucun ajustement n'était effectué.

- (2) S'il n'y a pas de distorsion imputable à la non-réponse, une corrélation entre X et R n'a pas d'effet (cas 4). Seule une corrélation entre X et Y réduit la variance (cas 5).

Faute d'avoir des données sur les non-répondants, il est impossible d'utiliser les données restantes pour trouver une variable auxiliaire X qui soit en forte corrélation avec Y. On peut cependant s'en servir pour rechercher des variables auxiliaires qui soient en forte corrélation avec le taux de réponse, R. Si l'on peut trouver une variable de ce genre, son utilisation pour la pondération par sous-groupes réduira la distorsion imputable à la non-réponse (lorsqu'elle existe), mais pas toujours la variance.

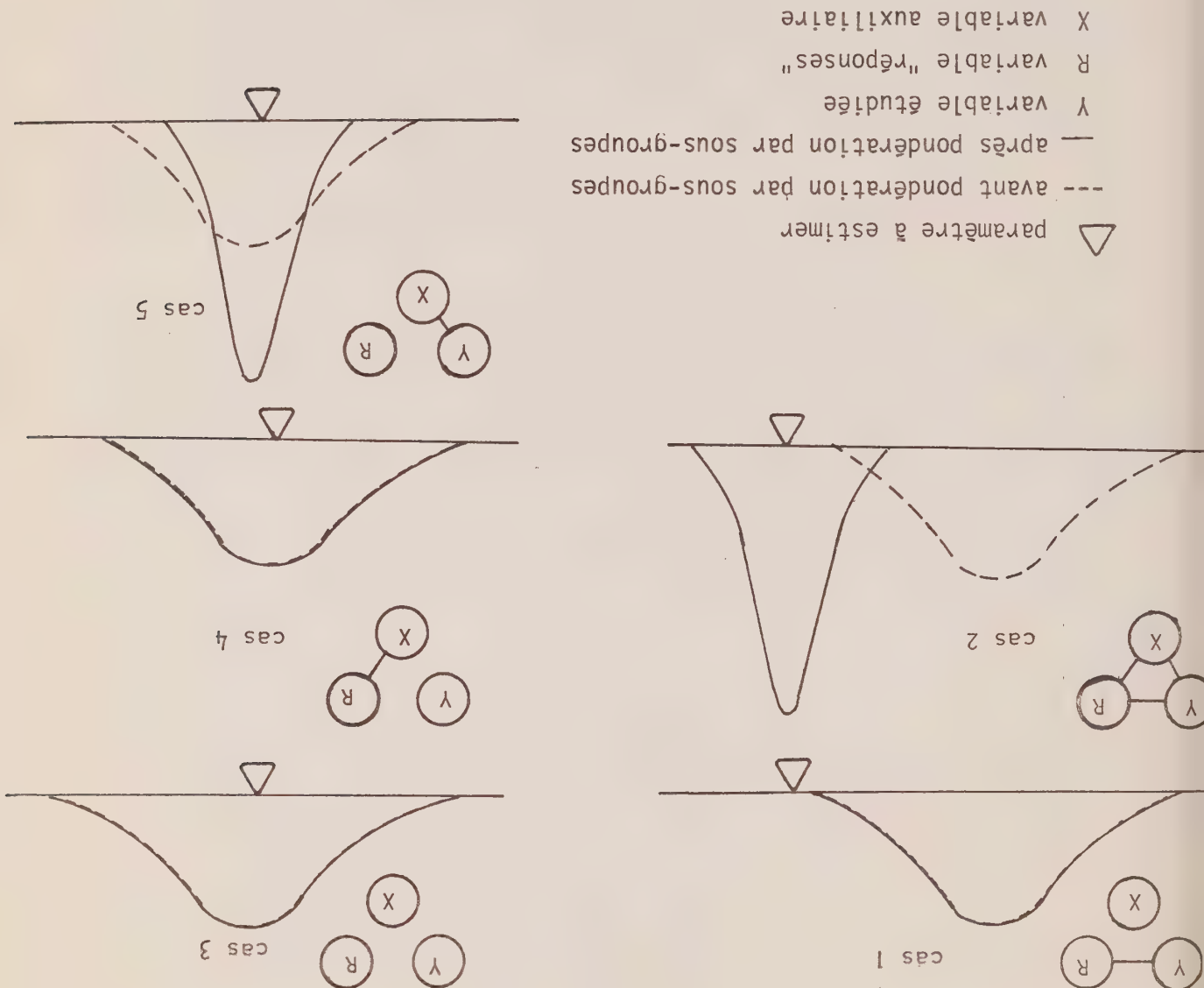
## 5. AUTRES METHODES D'AJUSTEMENT

Plusieurs autres méthodes d'ajustement sont exposées dans les ouvrages spécialisés. Il sera question de certaines d'entre elles dans la présente section. Il faut de plus amples recherches, dans le cas de telle ou telle d'entre elles, pour en dégager le véritable intérêt.

### 5.1. Pas d'ajustement

Il arrive qu'aucun ajustement ne soit nécessaire. Si l'on est certain qu'il n'existe aucune relation entre la variable étudiée et l'obtention de réponses, on peut considérer les réponses comme un échantillon aléatoire de la population. Dans le cas également où le résultat indiqué est réputé valoir uniquement pour la population de répondants potentiels, aucune correction n'est nécessaire. Dans toutes les autres situations, l'absence d'ajustement ne se justifie que si la catégorie "non-réponse" figure dans tous les tableaux publiés.

FIGURE 6. VARIANCE ET DISTORSION DES ESTIMATEURS AVANT ET APRES PONDERATION PAR SOUS-GROUPES



On peut en tirer un certain nombre de conclusions:

- (1) S'il existe une distorsion imputable à la non-réponse, la pondération par sous-groupes est significative lorsqu'il y a corrélation entre X et R (cas 2). La distorsion et la variance sont toutes deux réduites.

groupes et que  $X_2$  donne  $H$  groupes. En combinant  $X_1$  et  $X_2$  on obtient une subdivision en  $G \times H$  groupes. Si l'on ne connaît que les totaux marginaux  $N_{g+}^g (g=1, 2, \dots, G)$  de  $X_1$  et  $N_{+h}^h (h=1, 2, \dots, H)$  de  $X_2$ , on peut alors calculer de bonnes estimations  $\bar{N}_{gh}^g$  de  $N_{gh}^g$  en utilisant les renseignements que donne l'échantillon. Les poids sont alors égaux à :

$$\bar{w}_{gh} = \frac{N}{N_{g+}^g N_{+h}^h} \quad (g=1, 2, \dots, G; h=1, 2, \dots, H) \quad (9)$$

Quand on utilise ces trois estimateurs avec la même répartition en sous-groupes, ils ont tous la même distorsion, mais plus on dispose de renseignements au sujet de la dimension des sous-groupes, plus la variance de l'estimation sera faible. La pondération par sous-groupes a deux avantages : elle réduit la variance de l'estimation et elle réduit la distorsion imputable à la non-réponse. La figure 6 illustre les cas extrêmes. Si deux variables sont reliées entre elles, cela signifie qu'elles ont une forte corrélation.

$$\bar{y}_h = \frac{1}{m_h} \sum_{i=1}^m \bar{y}_{hi} \quad (h = 1, 2, \dots, H) \quad (5)$$

où  $\bar{y}_{h1}, \bar{y}_{h2}, \dots, \bar{y}_{hm_h}$  sont les valeurs de  $m_h$  éléments de sous-groupe  $h$  qui répondent. On combine ensuite les estimateurs des différents sous groupes,  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_H$  en un seul estimateur  $\bar{y}$  pour l'ensemble de la population.

$$\bar{y} = \frac{1}{H} \sum_{h=1}^H \bar{y}_h \quad (6)$$

Le type d'estimateur est déterminé par la quantité d'informations qui est disponible au sujet des poids  $\bar{w}_1, \bar{w}_2, \dots, \bar{w}_H$ .

Si l'on connaît les dimensions  $N_1, N_2, \dots, N_H$  des sous-groupes, la situation équivaut à une post-stratification (voir par exemple Holt & Smith, 1979). Les poids ne sont pas des quantités aléatoires, mais des quantités bien déterminées:

$$\bar{w}_h = \frac{N_h}{N} \quad (h = 1, 2, \dots, H) \quad (7)$$

Si ces dimensions ne sont pas connues, on peut les estimer en appliquant la formule

$$\bar{w}_h = \frac{n_h}{n} \quad (h = 1, 2, \dots, H) \quad (8)$$

dans laquelle  $n_h$  est le nombre d'éléments du sous-groupe  $h$  qui figurent dans l'échantillon ( $n = n_1 + n_2 + \dots + n_H$ ).

Dans une situation intermédiaire où l'on utilise deux variables auxiliaires  $X_1$  et  $X_2$  et où l'on ne connaît que les totaux marginaux des deux variables, on peut appliquer une autre méthode pour estimer les poids (voir par exemple Chapman, 1976). Supposons que  $X_1$  donne  $G$



#### 4. REDUCTION DE LA DISTORSION DUE A LA NON-REPONSE AU MOYEN

##### D'UNE PONDERATION PAR SOUS-GROUPES

Lorsque l'on constate ou que l'on soupçonne l'existence d'une relation entre la variable étudiée ( $Y$ ) et l'obtention de réponses ( $R$ ), il faut prendre des mesures pour réduire la distorsion imputable à la non-réponse. Il sera question dans la présente section d'un certain nombre de procédures d'ajustement qui sont fondées sur la pondération par sous-groupes. L'attention se portera essentiellement sur l'estimation de la moyenne de  $Y$  pour l'ensemble de la population.

On peut montrer que la distorsion introduite par le fait de n'utiliser que les valeurs tirées des réponses est proportionnelle à la covariance entre  $Y$  et  $R$ . S'il est possible de diviser la population en un certain nombre de sous-groupes pour lesquels la covariance est dans chaque cas négligeable, on peut alors combiner les estimations (pratiquement sans distorsion) des moyennes des différents sous-groupes en une estimation (pratiquement sans distorsion) de la moyenne pour l'ensemble de la population.

Considérons que la population finie se compose de  $N$  éléments  $U_1, U_2, \dots, U_N$  pour lesquels les valeurs de  $Y$  sont  $Y_1, Y_2, \dots, Y_N$ . Dans cette population on choisit, sans remise, un échantillon aléatoire simple  $u_1, u_2, \dots, u_n$  (les variables aléatoires sont souligées de dimension  $n$ ). Les valeurs correspondantes de  $Y$  sont  $Y_1, Y_2, \dots, Y_n$  et l'obtention ou la non-obtention d'une réponse est indiquée par  $r_1, r_2, \dots, r_n$  ( $r_i = 1$  indiquant qu'il y a une réponse et  $r_i = 0$  indiquant une non-réponse). En fait, on ne peut observer  $Y_i$  que dans le cas des éléments  $u_i$  de l'échantillon pour lesquels  $r_i = 1$ . Les  $m$  éléments qui répondent sont dénotés  $u_1^*, u_2^*, \dots, u_m^*$  ( $\bar{m} = r_1 + r_2 + \dots + r_n$ ), et pour ces éléments les valeurs de  $Y$  sont  $Y_1^*, Y_2^*, \dots, Y_m^*$ . Soit  $X$  une variable auxiliaire entraînant une division de la population en  $H$  sous-groupes de dimensions  $N_1, N_2, \dots, N_H$ . Dans la pondération par sous-groupes, on calcule tout d'abord dans chaque sous-groupe  $h$  un estimateur  $\bar{Y}_h^*$  de la

moyenne pour le sous-groupe :

Ce graphique contient environ 88% des informations existant au sujet des associations dans le tableau ( $\tau^2 = 0,88$ ). Les principales raisons qui expliquent la non-réponse, dans le cas des personnes âgées, sont le refus et la maladie. Dans le cas des jeunes, la non-réponse est due à l'impossibilité d'entrer en rapport avec les enquêtés : logement inhabité, absence du foyer et déménagement. On trouvera de plus amples détails sur l'application de l'analyse des correspondances dans Bethlehem et Kersten (1980).

### 3.3. Autres méthodes de sélection

Il y a de nombreuses autres méthodes, essentiellement non graphiques, qui servent à déterminer l'association existant entre les variables auxiliaires et l'obtention de réponses. On peut trouver dans Bishop, Fienberg & Holland (1975), par exemple, d'amples indications au sujet de l'association dans les tableaux de contingence.

Une méthode couramment utilisée pour la sélection des variables auxiliaires les plus importantes est la méthode AID (Automatic Interaction Detection), décrite par Morgan & Sonquist (1963). Par étapes successives on détermine les variables auxiliaires qui peuvent expliquer une aussi grande partie que possible de la variance caractérisant la variable binaire réponse/non-réponse. Il y a à cette méthode des inconvénients qui rendent sa fiabilité douteuse. Comme le processus de sélection se fait par étapes, on n'a aucune garantie de trouver la solution optimale. Du fait qu'il n'y a aucune règle fondée sur un modèle statistique qui indique à quel moment il faut arrêter le processus, le résultat est, en ce sens, également, assez arbitraire. De plus amples recherches dans ce domaine sont nécessaires; voir par exemple Kass (1980).

où  $\chi^2$  est la valeur de chi-carré pour le tableau et N le total général,

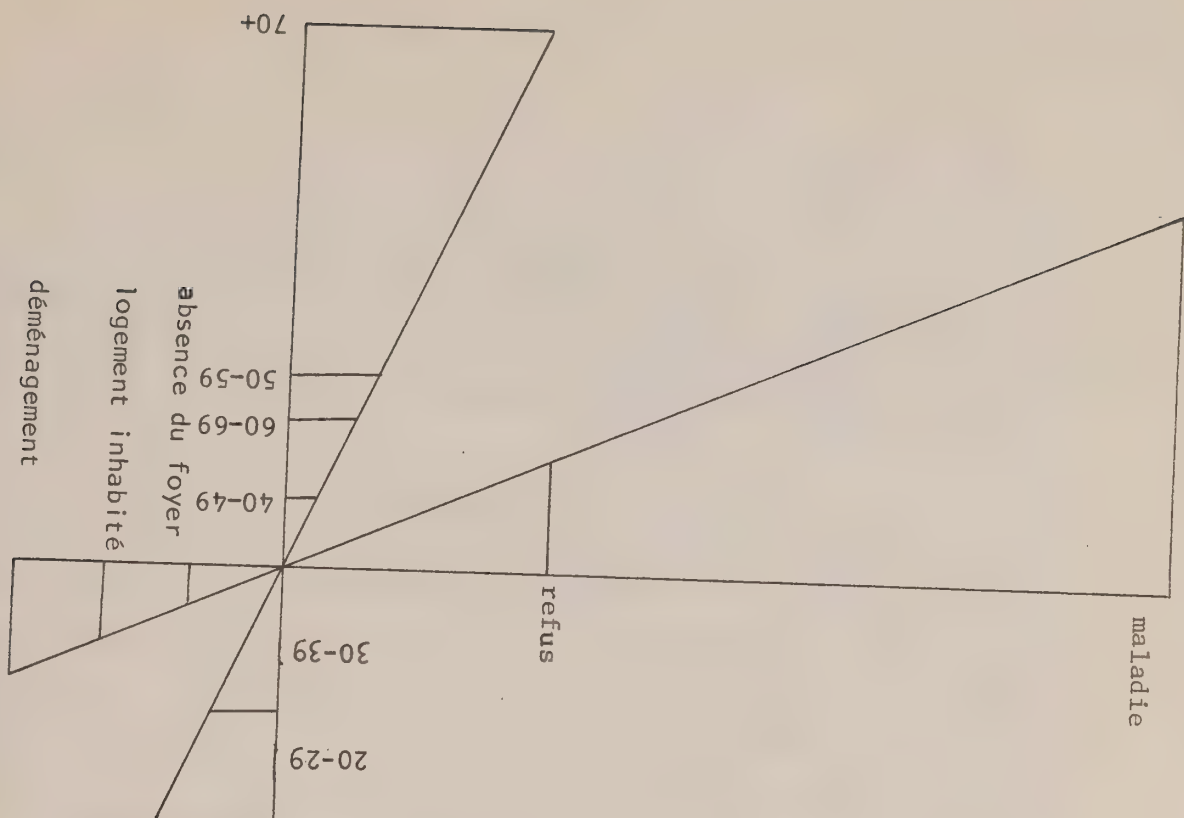
$$\tau_i^2 = N p_i^2 / \chi^2 \quad (4)$$

mesure la quantité d'information expliquée par le i-ème graphique

$$(i = 1, 2, \dots, s).$$

La figure 5 représente le premier diagramme en ailes de moulin établi pour deux variables: l'âge (six catégories) et le type de non-réponse (cinq catégories) au titre de l'Enquête de 1977/78 sur la demande de logements à Amsterdam.

FIGURE 5. DIAGRAMME EN AILES DE MOULIN INDICANT L'ASSOCIATION ENTRE L'ÂGE ET LE TYPE DE NON-REPONSE DANS L'ENQUÊTE DE 1977/78 SUR LA DEMANDE DE LOGEMENTS A AMSTERDAM



Il peut être bon de faire un certain nombre de remarques :

- (1) L'origine représente les deux distributions marginales du tableau.
- (2) Les valeurs réduites proches de l'origine représentent les catégories qui ressemblent à la distribution marginale et qui ont donc un comportement régulier. Les valeurs éloignées de l'origine représentent des catégories ayant un comportement différent.
- (3) La relation entre les deux variables est forte si les deux droites de régression sont voisines de la droite à coefficient angulaire de 45°.
- (4) La projection d'une variable appartenant à une catégorie à comportement différent sur l'axe de l'autre variable, par l'intermédiaire tant entre les diverses catégories dans lesquelles entrent les variables.

Le diagramme tel qu'il est décrit ci-dessus ne peut rendre compte de tous les renseignements figurant dans le tableau. Il les explique dans toute la mesure où cela est possible avec un graphique à deux dimensions. Sous certaines conditions on peut superposer au premier diagramme un second graphique qui rend compte autant que possible des renseignements non encore expliqués. Si besoin est on peut même en construire davantage, mais de préférence un seul suffit pour expliquer la plus grande partie des associations.

On peut établir au total  $s$  graphiques de ce genre,  $s$  étant inférieur d'une unité au nombre de lignes, ou au nombre de colonnes si ce dernier est moins élevé. Désignons par  $p_1, p_2, \dots, p_s$  les coefficients de corrélation maximisés. Etant donné que

$$\sum_{i=1}^s p_i^2 = X^2/N,$$

(3)

grille de points irrégulièrement espacés, qu'il n'est pas toujours facile d'interpréter. Pour simplifier l'interprétation, on porte sur le graphique des droites de régression au lieu des points eux-mêmes. En raison des propriétés particulières des valeurs réduites, la droite de régression expliquant les valeurs de  $y$  par rapport aux valeurs de  $x$ , dans le graphique, a la forme simple suivante:

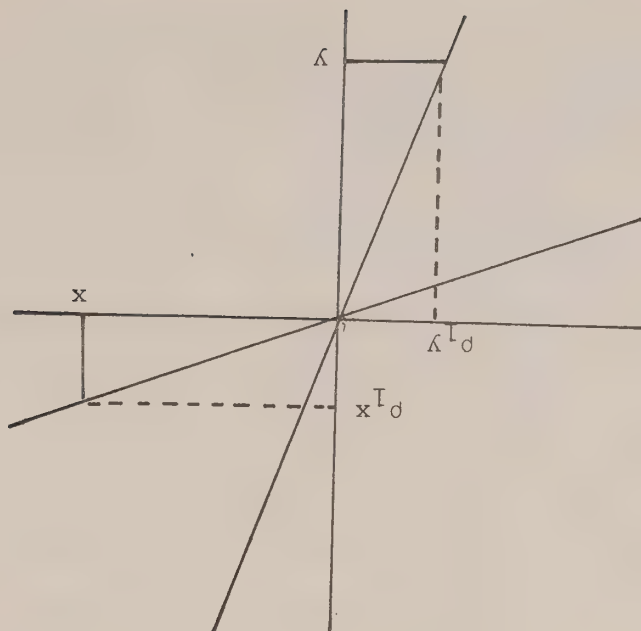
$$y = p_1 x \quad (1)$$

et la droite de régression expliquant les valeurs de  $x$  par rapport aux valeurs de  $y$  a la forme

$$x = p_1 y \quad (2)$$

$p_1$  étant le coefficient de corrélation maximisé. En portant sur le graphique les deux droites de régression, on obtient le diagramme en ailes de moulin - voir la figure 4.

FIGURE 4. LE DIAGRAMME EN AILES DE MOULIN





(2) L'ampleur de la non-réponse peut être évaluée d'après la distance entre les lignes de partage verticales et le côté droit du rectangle. Dans le présent exemple, il y a manifestement une quantité considérable de non-réponses.

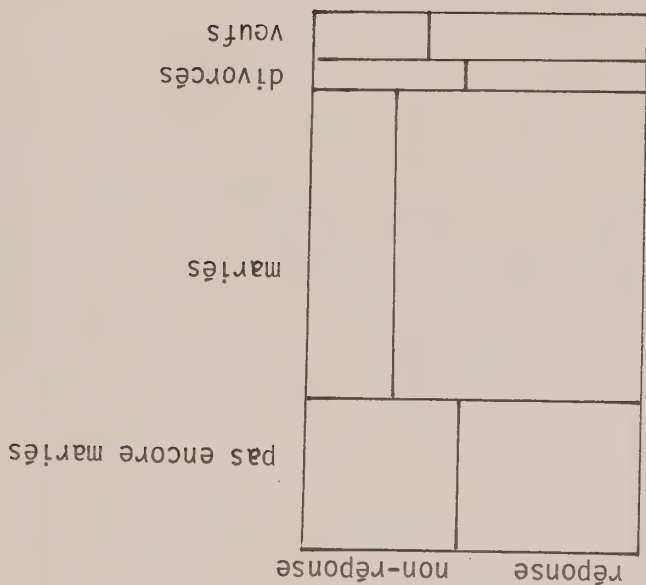
(3) Si toutes les lignes de partage constituent approximativement une droite, il n'y a pas de relation entre le taux de réponse et la variable auxiliaire. Dans le présent cas, il y a manifestement une relation: les personnes mariées répondent mieux que les autres. Le taux de réponse est plus faible dans le groupe des personnes non encore mariées et dans celui des divorcés.

On trouvera de plus amples détails au sujet du diagramme à cases dans Bethlehem et Kersten (1981).

### 3.2.2. Le diagramme en ailes de moulin

Le diagramme en ailes de moulin est une représentation graphique des résultats de l'analyse des correspondances. L'analyse des correspondances est une technique d'analyse des associations existant dans les tableaux à double entrée (voir par exemple Benzécri, 1976). Les lignes du tableau (catégories dans lesquelles entre la variable faisant l'objet d'une tabulation verticale) et les colonnes (catégories dans lesquelles entre la variable faisant l'objet d'une tabulation horizontale) sont représentées géométriquement. Cette représentation géométrique contient toutes les informations disponibles au sujet des associations existant dans le tableau. Par un système de mise à échelle, on affecte aux lignes et aux colonnes des valeurs réduites telles que le coefficient de corrélation calculé au moyen de ces valeurs soit porté à son maximum. A chacune des cases du tableau correspondent deux valeurs réduites: une pour la ligne et une pour la colonne. Si l'on conçoit ces valeurs comme des coordonnées, on peut établir un graphique correspondant au tableau. Cela donne une

Figure 3. DIAGRAMME A CASES INDICANT LA SITUATION MATRIMONIALE DES ENQUETES LORS DE L'ENQUETE DE 1977/78 SUR LA DEMANDE DE LOGEMENTS A AMSTERDAM



Il peut être bon d'appeler l'attention sur un certain nombre de points: (1) La hauteur des bandes indique dans quelle mesure les diverses catégories contribuent à l'échantillon. De toute évidence, une forte proportion de la population est mariée. La catégorie la plus petite est celle des personnes divorcées.

### 3.2.1. Le diagramme à cases

Le diagramme à cases peut être considéré comme une généralisation de l'histogramme ou du diagramme en bâtons. Ce nom lui a été donné à cause de sa forme (voir la figure 2).

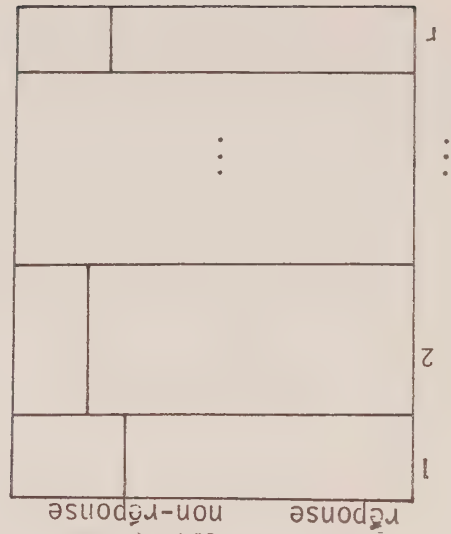
Un rectangle de largeur standard et de hauteur proportionnelle à la taille de l'échantillon représente l'échantillon. Le rectangle est divisé en un certain nombre de bandes (correspondant aux catégories dans lesquelles peut entrer la variable auxiliaire). La hauteur

d'une bande particulière est proportionnelle au nombre d'éléments de l'échantillon qui figurent dans la catégorie correspondante. Chaque bande est divisée par une ligne verticale en une case gauche (la

réponse) et une case droite (la non-réponse). La grandeur de ces deux cases est proportionnelle au nombre de réponses et de non-réponses, respectivement dans la catégorie considérée. La figure 3 donne un

exemple de diagrammes à cases. Les données proviennent de l'enquête de 1977/78 sur la demande de logements à Amsterdam. La variable auxiliaire est la situation matrimoniale de la personne figurant dans l'échantillon.

FIGURE 2. LE DIAGRAMME A CASES



Il est admis que les variables auxiliaires sont des variables nominales, c'est-à-dire que des valeurs différentes de ces variables servent uniquement à faire une distinction entre différents groupes. Il n'est pas permis de faire entrer ces valeurs, qui ne sont en fait que des étiquettes dans des opérations arithmétiques.

L'hypothèse selon laquelle les variables sont nominales n'est en pratique pas une restriction. De nombreuses variables sont nominales et d'autres types de variables peuvent facilement être réexprimées sous forme de variables nominales. A titre d'exemple illustrant la quantité de renseignements auxiliaires disponibles, nous donnons ci-après la liste des variables auxiliaires utilisées dans l'enquête de 1977/78 sur la demande de logements :

1) année de naissance	7) nombre d'étages dans le logement
2) sexe	8) année de construction du logement
3) situation matrimoniale	9) municipalité
4) dimension de la famille	10) quartier de la ville
5) structure de la famille	11) degré d'urbanisation
6) type de logement	

### 3.2 Méthodes graphiques

Comme instrument préliminaire pour la sélection des variables auxiliaires, on a mis au point des méthodes graphiques. L'avantage de ces méthodes est évident. Elles mettent en lumière des faits et des relations qui ne sont pas immédiatement visibles et elles peuvent stimuler et faciliter l'analyse. Elles permettent souvent de comprendre plus complètement et d'une manière mieux équilibrée que des tableaux ou des textes explicatifs. En outre, les relations qui s'en dégagent sautent plus clairement aux yeux et sont plus facilement inscrites dans la mémoire (voir Schmid, 1954). Deux méthodes graphiques simples sont présentées dans les sections qui suivent : le diagramme à cases et le diagramme en ailes de moulin.

### 3. CHOIX DES VARIABLES AUXILIAIRES

#### 3.1 Variables auxiliaires

Il importe de découvrir s'il existe éventuellement une relation entre la variable étudiée et l'obtention de réponses. Il n'est cependant pas possible de dégager cette relation au moyen des données de sondage, car les valeurs de la variable étudiée ne sont pas connues dans le cas des non-répondants. Pour pouvoir dire quelque chose au sujet des non-répondants, on doit avoir des renseignements à leur sujet. Une source d'information concernant les non-réponses est constituée par les variables auxiliaires. Il s'agit de variables que l'on peut mesurer pour les répondants et pour les non-répondants. Il existe deux types de renseignements auxiliaires :

(1) Les renseignements que l'enquêteur peut recueillir sans procéder à un interrogatoire direct, comme par exemple le type de ville, le type de logement, l'année (approximative) de construction du logement et le statut social des personnes vivant dans le voisinage;

(2) Les renseignements qu'il est possible d'obtenir dans les archives administratives, concernant notamment l'âge, le sexe et la situation matrimoniale.

L'analyse de la relation entre les variables auxiliaires et le taux de réponse jette quelque lumière sur le groupe des personnes qui ne répondent pas. Elle peut renseigner aussi sur la relation existant entre la variable étudiée et le taux de réponse. Les variables auxiliaires qui font ressortir une relation nette avec le taux de réponse jouent un rôle important dans les procédures d'ajustement, ainsi qu'il en sera question plus loin.



ristiques de l'enquête (but, type de questions, techniques d'interrogatoire, enquêteurs, période où s'effectue le travail sur le terrain, etc.). On choisit l'échantillon parmi la population sans tenir compte de ces deux strates. Par conséquent, le nombre de répondants est une variable aléatoire dans l'un et l'autre modèles.

Si, au lieu d'un sondage, on procède à un dénombrement complet, la détermination des répondants reste alors un processus aléatoire dans le cas du modèle à taux de réponse aléatoire, tandis qu'elle serait un processus fixe dans le cas de l'autre modèle. Il y a cependant une certaine ressemblance entre les deux modèles. Si l'on suppose qu'il existe deux mécanismes stochastiques, le mécanisme de l'échantillonnage et le mécanisme de la réponse, les deux modèles ne diffèrent que par l'ordre dans lequel les mécanismes sont appliqués : dans le modèle à taux de réponse fixe, c'est d'abord le mécanisme de la réponse qui entre en jeu pour chaque élément de la population. Cela détermine les deux strates. L'échantillon est choisi ensuite. Dans le modèle à taux de réponse aléatoire, on choisit d'abord l'échantillon, puis le mécanisme de la réponse entre en jeu pour chacun des éléments choisis. Le point de savoir quel modèle il faut utiliser est plus ou moins une question de préférence personnelle. Le modèle à taux de réponse aléatoire donne la possibilité d'estimer les probabilités de réponse. On peut se servir de ces probabilités dans les procédures d'ajustement, ou bien les rattacher à des caractéristiques personnelles. Le modèle à taux de réponse fixe aboutit généralement à des formules plus simples. La théorie fondée sur ce modèle est conditionnée par la composition effective des strates de répondants et de non-répondants. Par conséquent, on peut calculer le degré d'exactitude des estimations, mais non déterminer l'exactitude de la méthode d'estimation. A cause de ce dernier argument, les recherches portent essentiellement sur le modèle à taux de réponse aléatoire.

en question sont probablement en train de passer leur temps de loisir quelque part ailleurs. La même chose vaut pour l'enquête sur le nombre d'heures que les gens passent à regarder la télévision: les absents du foyer (le soir) ne sont probablement pas en train de regarder la télévision. L'un des objectifs de l'enquête sur la demande de logements est de mesurer la fréquence avec laquelle les gens déménagent. Comme il y a une quantité considérable de non-réponse dues aux déménagements (l'unité d'échantillonnage est une personne), l'estimation concernant la population totale sera entachée de distorsion. Un certain nombre d'enquêtes montrent que les personnes non mariées ont un taux de réponse plus faible. S'il existe une relation entre la situation matrimoniale et la variable étudiée, alors les estimations seront faussées dans ce cas également.

## 2.3 Modèles relatifs à l'obtention de réponses

La première chose à faire, lorsqu'on veut mettre au point des théories pour le traitement de la non-réponse, est d'élaborer un modèle mathématique qui décrit la manière dont fonctionne le mécanisme de la non-réponse. Il est souvent question dans les ouvrages techniques de deux modèles de ce genre, que nous appellerons ici le "modèle à taux de réponse aléatoire" et le modèle "à taux de réponse fixe".

D'après le modèle à taux de réponse aléatoire, chaque élément de la population est caractérisé par une certaine probabilité (inconnue) de réponse. Ces probabilités de réponse ne sont pas nécessairement les mêmes pour tous les éléments. Lorsque l'enquêteur se met en rapport avec la personne à questionner, le mécanisme des probabilités entre alors en jeu, ce qui détermine si la personne répondra ou non. Le modèle à taux de réponse fixe suppose l'existence de deux strates dans la population: une strate de répondants potentiels et une strate de non-répondants potentiels. La dimension et la composition de chaque strate ne sont pas connues d'avance. Elles sont déterminées par les caracté-

Tableau 2  
Pourcentages de non-réponse dans certaines enquêtes du BSC

Année	EMO	ESC	ECV	ENV	EV
	tn rn	tn rn	tn rn	tn rn	tn rn
1973	13.2				
1974		28.2	15.6		
1975	15.3	30.1	18.3		14.5
1976		28.1	18.6	23.0 <sup>1)</sup>	12.9
1977	13.1	6.6	30.9	20.5	17.6
1978		36.1	23.9		9.3
1979	19.7	36.6	24.4	33.7 <sup>2)</sup>	25.5
1980		36.8	24.7	35.6	24.5
				19.7	
				31.1	
				30.6	
				23.9	
				26.2	
				21.9	
				12.5	

1) personnes agréées seulement

2) personnes jeunes seulement

tn = pourcentage de non-réponses toutes catégories

rn = pourcentage de refus

EMO = Enquête sur la main-d'oeuvre

ESC = Enquête sur les sentiments des consommateurs

ECV = Enquête sur les conditions de vie

ENV = Enquête nationale sur les voyages

EV = Enquête sur les vacances

Ainsi qu'il a été mentionné plus haut, l'existence d'une relation entre la variable étudiée et l'obtention de réponses réduit la valeur des conclusions de l'enquête. Il n'est pas rare qu'une telle relation existe, comme le montreront les exemples ci-après. Si l'enquête a pour objet de déterminer de quelle manière les gens occupent leurs loisirs, la raison des non-réponses imputables à une "absence du foyer" est alors assez difficile à déterminer, du fait que les personnes

maieur sur le plan pratique. La nature des bases de sondage disponibles est une considération importante dans la constitution de l'échantillon. Parmi les facteurs qui entrent en jeu il y a le type d'unité d'échantillonnage, le taux de couverture, l'exactitude et l'exhaustivité de la liste, ainsi que la quantité de renseignements auxiliaires figurant dans la liste et leur qualité.

Pour les bases de sondage dans lesquelles l'unité d'échantillonnage est une personne, le BCS ne peut disposer que des registres administratifs des autorités locales (municipalités). Pour les enquêtes sur les ménages il peut établir sa propre base de sondage, mais pour le moment il juge approprié d'utiliser la liste des points de distribution de l'administration postale.

## 2.2 L'ampleur de la non-réponse

Il est assez difficile de comparer les taux de non-réponse enregistrés à l'occasion de différentes enquêtes. Le taux de non-réponse dépend d'un certain nombre de circonstances: but de l'enquête, type d'unité d'échantillonnage, plan de sondage, efficacité des travaux effectués sur le terrain, efficacité des enquêteurs, mesures visant à réduire la non-réponse, période au cours de laquelle se déroule l'enquête, population prise comme objectif, longueur du questionnaire, libellé des questions, etc. Même la définition de la non-réponse peut varier. Il est nécessaire d'établir un cadre qui permette de comparer correctement différentes enquêtes. En tenant dûment compte des facteurs qui influent sur le nombre de non-réponses, on peut juger de la qualité des différentes enquêtes. Un tel cadre offre également la possibilité de comparer des enquêtes effectuées dans des pays différents.

Le tableau 2 présente les pourcentages de non-réponse enregistrés à l'occasion d'un certain nombre d'enquêtes effectuées par le BCS. On peut y constater que ce pourcentage manifeste une nette tendance à augmenter.



(5)

Renseignements égarés. Des renseignements peuvent se perdre

après une enquête sur le terrain. Certains questionnaires sont parfois inutilisables à cause de leur mauvaise qualité ou parce que le répondant a triché. Il arrive aussi que d'autres aient été perdus ou oubliés.

La typologie décrite ci-dessus est applicable dans la plupart des types d'enquête, mais il faut prendre des précautions dans le cas des plans de sondage complexes. Lorsqu'on effectue par exemple un sondage à plusieurs degrés, cette typologie peut être utilisée à chacun des degrés. On peut classer la même source d'erreur différemment aux différents degrés. Donnons un exemple à titre d'illustration: dans une enquête sur les ménages, un échantillon de ménages est d'abord choisi; l'enquêteur dresse la liste de toutes les personnes faisant partie des ménages ainsi choisis et, à partir de cette liste, sélectionne un échantillon. Lors d'une telle énumération l'étudiant qui vit dans une chambre de bonne est souvent "oublié". Au premier degré de la procédure de sondage, cette situation serait considérée comme une erreur de mesure, et au deuxième degré comme une erreur par défaut de couverture.

Pour certaines sources d'erreur, la classification peut dépendre d'autres facteurs. Si une personne appelée à être interrogée meurt avant que l'entrevue ne puisse avoir lieu, la classification dépend de la date du décès. Lorsque le décès est intervenu avant le jour où l'on a choisi l'échantillon, on a alors un taux de couverture excessif, mais s'il s'est produit entre le jour où l'échantillon a été constitué et le jour de l'entrevue, il s'agit alors en fait d'une non-réponse.

Avant de choisir l'échantillon, il faut subdiviser la population en différentes parties que l'on appelle les unités d'échantillonnage. A chaque élément de la population doit correspondre une unité d'échantillonnage et une seule. L'établissement de la liste d'unités d'échantillonnage, appelée base de sondage, soulève souvent un problème



(2) Réfus. Certaines des causes de refus sont temporaires et

peuvent disparaître. Il est possible qu'une personne refuse parce qu'elle est mal disposée ou qu'on la contacte à une heure qui ne lui convient pas. Elle peut fort bien se montrer coopérative lors d'un autre essai ou d'une autre visite. Etant donné cependant qu'un nombre assez important de refus peuvent être considérés comme définitifs, il vaut mieux dans ce cas parler de réponses impossibles à obtenir, pour montrer qu'il s'agit d'un refus catégorique plutôt qu'un retard dans l'observation. Des tentatives répétées n'auraient aucun succès. De ce point de vue on classe dans cette catégorie, plutôt que dans celle des absents du foyer, les répondants que l'on sait être absents pendant toute la période de l'enquête.

(3)

Incapacité. Ce terme peut être utilisé dans le cas où la non-réponse est due à une maladie mentale ou physique qui empêche l'enquête de répondre pendant toute la période de l'enquête. On range aussi dans cette catégorie les non-réponses dues à la méconnaissance de la langue. En généralisant on pourrait grouper cette catégorie avec celle, définie ci-dessus, des non-réponses dues à l'impossibilité d'obtenir des renseignements. Il peut cependant être utile, dans certains cas, de faire la distinction entre les enquêtes qui ne veulent pas répondre et ceux qui le voudraient, mais qui en sont incapables.

(4)

Introuvables. Cette catégorie peut être importante, par exemple dans le cas des personnes qui se déplacent constamment. On s'abstient alors de les identifier ou de les suivre parce que cela serait trop coûteux. Dans cette même catégorie générale sont rangés les cas où aucune tentative d'entrevue n'est faite, par exemple pour des raisons d'inaccessibilité (gardien de phare, berger) ou de danger possible (chien de garde, taudis).

Par défaut de couverture on entend toutes les erreurs qui résultent des différences existant entre la population-objectif et la population sondée. Les éléments qui font partie à la fois de la population-objectif et de la population sondée sont les éléments "corrects". Lorsque des éléments de la population-objectif ne figurent pas dans la population sondée, on a un taux de couverture réduit. Ces éléments n'ont aucune chance (probabilité nulle) de figurer dans l'échantillon. Lorsque des éléments de la population sondée ne font pas partie de la population-objectif, on a un taux de couverture excessif (comptages ou observations en trop). Il faut exclure ces éléments de l'échantillon avant de procéder à l'analyse. Si cet excès de couverture est inattendu, la taille de l'échantillon définitif peut être inférieure à celle de l'échantillon prévu.

La non-réponse est le fait de ne pas obtenir d'observations au sujet de certains éléments choisis et désignés pour faire partie de l'échantillon. Une bonne classification des erreurs dues à la non-réponse dépend des conditions dans lesquelles se déroule l'enquête. La classification donnée ci-après est axée essentiellement sur les problèmes rencontrés lors des interrogatoires directs. Un traitement analogue peut être appliqué dans d'autres cas où l'enquête se déroule de manière différente. On peut distinguer entre les diverses catégories ci-après de non réponses:

- (1) Absents du foyer. Pour réduire l'ampleur de cette catégorie, on peut procéder à des "rappels". Il y a lieu de faire des recherches sur le nombre optimal de rappels. Il serait utile d'utiliser pour cette catégorie l'expression temporairement indisponible, signifiant que l'entrevue est retardée plutôt que refusée. Le répondant peut être trop occupé, fatigué ou malade lors de la première entrevue, mais il se montrera coopératif à l'occasion d'une nouvelle visite.

Les erreurs d'observation sont dues à ce que certaines observations sont obtenues et enregistrées de manière incorrecte. On peut les subdiviser encore en erreurs de mesure et erreurs de traitement.

Les erreurs de mesure sont dues soit à l'enquêteur, soit au répondant. L'enquêteur lui-même peut être une source d'erreur. Il peut influencer la réponse par sa seule présence, par le fait qu'il appartient à tel ou tel des deux sexes, par sa peau, sa couleur, son âge ou sa façon de s'habiller. La manière dont il pose les questions et explique les réponses affecte également les résultats. La réponse fournie par une personne peut dépendre du type de question (selon qu'il s'agit d'une question mesurant un fait, comme l'année de naissance, ou une opinion). Des erreurs peuvent aussi être introduites par d'autres facteurs, selon par exemple que la personne interrogée comprend ou non la question, qu'elle connaît ou non la réponse, qu'elle tient à garder la réponse pour elle-même, ou qu'elle désire donner d'elle une certaine image. En outre, la mémoire n'est pas toujours exempte d'erreurs.

Les erreurs de traitement se produisent lors du traitement des données par le service statistique, lors du codage, de la mise en tableaux et des calculs statistiques.

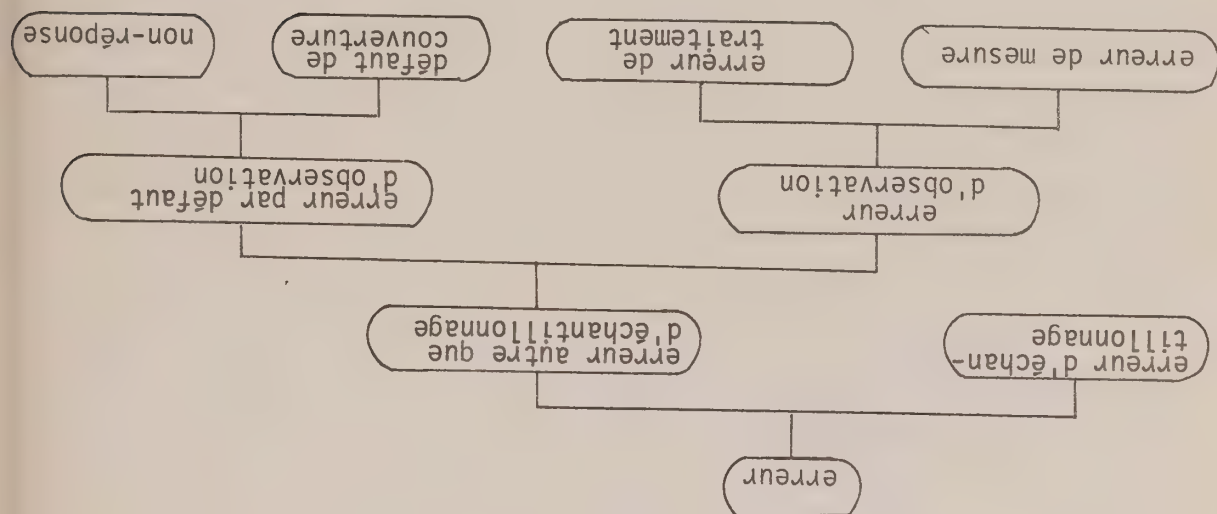
Les erreurs par défaut d'observation sont dues à l'impossibilité d'obtenir des observations concernant certaines parties de la population. On peut les subdiviser en erreurs dues à un défaut de couverture et à la non-réponse.

Appelons population-objectif la population sur laquelle l'enquête est censée porter. Des difficultés pratiques qui se présentent dans le cas de certaines parties de la population peuvent mener à les éliminer du champ de l'enquête. Il est possible aussi que la population effectivement sondée contienne des éléments qui ne font pas partie du champ de l'enquête.

Les deux sources d'erreur, dans les enquêtes, sont les erreurs d'échantillonnage et les erreurs autres que d'échantillonnage.

Les erreurs d'échantillonnage sont constituées par la partie de l'erreur qui est due au fait qu'on observe uniquement un échantillon de valeurs, et non la population totale. L'erreur d'échantillonnage a une distribution de fréquences probables qui est constituée par la totalité des erreurs d'échantillonnage de tous les échantillons possibles de même dimension. On se sert de cette distribution pour estimer la caractéristique de la population que l'on veut étudier.

FIGURE 1. TYPOLOGIE DES ERREURS QUI AFFECTENT LES ENQUÊTES



Les erreurs autres que d'échantillonnage sont celles que, dans les estimations portant sur un échantillon, on ne peut attribuer aux fluctuations de l'échantillonnage. Elles posent souvent un problème plus sérieux que les erreurs d'échantillonnage. On peut les répartir en erreurs d'observation et erreurs par défaut d'observation.



Les sections ci-après donnent un aperçu des travaux d'analyse effectués par le BCS au sujet des non-réponses. La section 2 présente des définitions, ainsi que les problèmes relatifs. Elle donne des chiffres concernant le taux de non-réponse enregistré à l'occasion d'enquêtes effectuées par le BCS. Dans la section 3 il sera question de méthodes graphiques utilisées pour choisir des variables auxiliaires. Ces méthodes jettent quelque lumière sur la non-réponse et peuvent être utilisées dans les procédures d'ajustement. La section 4 présente les méthodes d'ajustement fondées sur la pondération par sous-groupes et la section 5 donne des indications sur un certain nombre d'autres méthodes.

## 2. LE PHENOMENE DE LA NON-REPONSE

Dans la présente section le problème de la non-réponse est replacé dans un cadre général, où un certain nombre d'autres problèmes propres aux enquêtes par sondage jouent également un rôle. Des chiffres sont donnés au sujet des non-réponses constatées à l'occasion d'enquêtes effectuées par le BCS. Il sera question aussi de certains cas où une relation existe entre la variable étudiée et l'obtention de réponses. Dans la dernière partie de cette section seront examinés deux modèles concernant le mécanisme de la non-réponse.

### 2.1 Terminologie

L'objectif de toute enquête est de déterminer certaines caractéristiques de la population. Par suite de toutes sortes d'erreurs, la véritable valeur de ces caractéristiques ne sera généralement jamais obtenue. Une typologie des sources d'erreur est présentée dans la figure 1, qui est un schéma dû à Kish (1967).



renseignements qui est obtenue, les estimations relatives aux paramètres démographiques sont moins précises. En second lieu, s'il existe une relation entre la variable étudiée et l'obtention de réponses, les calculs effectués sur la base des réponses ne sont pas valables pour l'ensemble de la population. Par exemple, si la demande de logements des répondants est supérieure à celle des non-répondants, les estimations de la demande de logements dans l'ensemble de la population seront sensiblement trop élevées.

Il est évident que le taux de non-réponse doit être maintenu à un niveau aussi bas que possible. Si, malgré ces efforts, il subsiste encore une quantité considérable de non-réponses, il faut alors prendre des dispositions pour éviter de formuler des conclusions erronées au sujet de la population. La combinaison de procédures d'ajustement et des techniques habituelles d'estimation doit aboutir à des estimations valables concernant la population.

Deux départements du BCS (Bureau central de statistique des Pays-Bas) participent aux recherches sur la non-réponse. Le Département des enquêtes sociales est responsable du travail effectué sur le terrain à l'occasion des enquêtes. Il s'attache à supprimer les non-réponses lors de la collecte des données. Il effectue des recherches sur le nombre optimal de rappels et le meilleur moment pour l'entrevue (voir Widdershoven et Van den Berg (1980)). Il procède à des expériences pour trouver la meilleure manière d'entrer en contact avec les personnes et les ménages au moyen de lettres introductives. Il s'efforce de mesurer l'impact de la lassitude engendrée par des entrevues trop longues ou trop fréquentes. En fin de compte, malgré ces efforts, il y a toujours une certaine quantité de non-réponses. Le Département des méthodes statistiques entreprend des recherches au sujet de l'effet des non-réponses sur l'exactitude des résultats de l'enquête. Il met au point des méthodes permettant d'ajuster les estimations démographiques pour éliminer la distorsion imputable aux non-réponses. La suite du présent document est essentiellement consacrée aux travaux de ce dernier département.

## LE PROBLEME DE LA NON-REPONSE

J.G. Bethlehem et H.M.P. Kersten

Ce document donne un aperçu des recherches effectuées au sujet des non-réponses au Bureau central de statistique (BCS) des Pays-Bas. Le phénomène de la non-réponse est replacé dans un cadre général. Des indications sont données, au sujet de l'ampleur de la non-réponse, au moyen de chiffres tirés d'un certain nombre d'enquêtes effectuées par le BCS. Il est question de l'utilisation de variables auxiliaires comme moyen d'obtenir des renseignements au sujet des non-répondants. Ces variables peuvent soit servir à dégager les caractéristiques des non-répondants, soit être utilisées pour une stratification dans les procédures d'ajustement.

Il est question en plus grand détail de l'ajustement destiné à éliminer la distorsion imputable aux non-réponses au moyen d'une pondération par sous-groupes. Enfin, la dernière section énumère un certain nombre d'autres méthodes qui visent également à réduire cette distorsion.

## 1. INTRODUCTION

La non-réponse devient depuis quelque temps un sujet de préoccupation croissante dans la recherche relative aux enquêtes. Le phénomène de la non-réponse, qui est dû à ce que certaines personnes ne sont pas en mesure ou ne sont pas désireuses de répondre aux questions posées par l'enquêteur, peut apparaître dans les enquêtes par sondage aussi bien que dans les recensements. Il affecte la qualité de l'enquête de diverses manières. Tout d'abord, par suite de la réduction de la quantité de

J.G. Bethlehem et H.M.P. Kersten, Bureau central de statistique des Pays-Bas.  
Les vues exprimées dans ce document sont celles de l'auteur et ne traduisent pas nécessairement la politique du Bureau central de statistique des Pays-Bas.

- [47] Sedransk, S. and Meyer, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling", J. Roy. Statist. Soc. B., 40, 239-252.
- [48] Scott, A. and Holt, D. (1981), "The effect of Two-Stage Sampling on Ordinary Least Squares Methods", (non publié).
- [49] Shah, B.V. (1978), "SUDAN: Survey Data Analysis Software", Proc. Statist. Comp. Sect., Amer. Statist. Assoc., 146-151.
- [50] Shah, B.V., Holt, M.M. and Folsom, R.E. (1977), "Inference About Regression Models from Sample Survey Data", Bull. Inter. Statist. Inst. 47, Bk. 3, 43-57.
- [51] Shuster, J.J. and Downing, D.J. (1976), "Two-Way Contingency Tables for Complex Sampling Schemes", Biometrika 63, 271-278.
- [52] Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review (Avec Commentaires), J. Roy. Statist. Soc. A., 139, 183-195.
- [53] Tepping, B.J. (1968), "The Estimation of Variance in Complex Surveys", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 66, 411-414.
- [54] Thomsen, I. (1978), "Design and Estimation Problems When Estimating a Regression Coefficient from Survey Data", Metrika 25, 27-35.
- [55] Tomberlin, T.J. (1979), "The Analysis of Contingency Tables of Data from Complex Samples", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 152-157.
- [56] Woodruff, Ralph S. (1952), "Confidence Intervals for Medians and Other Position Measures", J. Amer. Statist. Assoc. 47, 635-646.

[38] Nathan, G. (1973), "Approximate Tests of Independence in Contingency Tables from Complex Stratified Samples", National Center for Health Statistics, Vital and Health Statistics Series 2, No. 53, Washington, D.C.

[39] Nathan, G. (1975), "Tests of Independence in Contingency Tables from Stratified Proportional Samples", Sankhya C, 37, 77-87. [erratum: Sankhya C, 40, (1978), 190].

[40] Nathan, G. and Holt, D. (1980), "The Effect of Survey Design on Regression Analysis", J. Roy. Statist. Soc. B, 42, 377-386.

[41] Pfeffermann, D., and Nathan, G. (1981), "Regression Analysis of Data from Complex Samples", J. Amer. Statist. Assoc. 76, 681-689.

[42] Porter, R.M. (1973), "On the Use of Survey Sample Weights in the Linear Model", Annals of Economic and Social Measurement, 2, 141-158.

[43] Rao, J.N.K. (1975), "Analytic Studies of Sample Survey Data", Survey Methodology, Vol. 1, Supplementary Issue.

[44] Rao, J.N.K. and Scott, A.J. (1981), "The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", J. Amer. Statist. Assoc. 76, 221-230.

[45] Richards, V. and Freeman, D.H. Jr. (1980), "A Comparison of Replicated and Pseudo-Replicated Covariance Matrix Estimators for the Analysis of Contingency Tables", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 209-211.

[46] Särndal, C.E. (1978), "Design-Based and Model-Based Inference in Survey Sampling", Scand. J. Statist., 5, 27-52.



[128] Kish, L. and Frankel, M.R. (1974), "Inference from Complex Samples", (avec commentaires), J. Roy. Statist. Soc. B, 36, 1-37.

[129] Koch, G.G., Freeman, D.J., Jr., and Freeman, J.L. (1975), "Strategies in The Multivariate Analysis of Data from Complex Surveys", Inter. Statist. Rev. 43, 59-78.

[130] Koch, G.G., Stokes, M.E. and Brock, D. (1980), "Applications of Weighted Least Squares Methods for Fitting Variational Models to Health Survey Data", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 218-223.

[131] Konijn, H.S. (1962), "Regression Analysis for Sample Surveys", J. Amer. Statist. Assoc. 57, 590-606.

[132] Krewski, D., and Rao, J.N.K. (1981), "Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods", Ann. Statist., 9 (5/ 1010-1019).

[133] Lepkowski, J.M. and Landis, J.R. (1980), "Design Effects for Linear Contrasts of Proportions and Logits", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 224-229.

[134] McCarthy, P.J. (1969), "Pseudo-Replication: Half-Samples", Inter. Statist. Rev. 37, 239-264.

[135] Miller, R.G. (1974), "The Jackknife--A Review", Biometrika 61, 1-15.

[136] Nathan, G. (1969), "Tests of Independence in Contingency Tables from Stratified Samples", New Developments in Survey Sampling (N.L. Johnson and H. Smith, eds.), New York: Wiley, 578-600.

[137] Nathan, G. (1972), "On the Asymptotic Power of Tests for Independence in Contingency Tables from Stratified Samples", J. Amer. Statist. Assoc., 67, 917-920.



- [18] Hidiroglou, M.A. and Rao, J.N.K. (1981), "Chi-Square Tests for the Analysis of Categorical Data from the Canada Health Survey", Invited Paper for 43rd Session of I.S.I., Buenos-Aires.
- [19] Holt, D., Richardson, S.C. and Mitchell, P.W. (1980), "The Analysis of Correlations in Complex Survey Data", (non publié).
- [20] Holt, D. and Scott, A.J. (1981), "Regression Analysis Using Survey Data", The Statistician, 30. (à paraître).
- [21] Holt, D., Scott, A.J., and Ewings, P.O. (1980), "Chi-Squared Test with Survey Data", J. Roy. Statist. Soc. A, 143, 302-330.
- [22] Holt, D. and Smith, T.M.F. (1979), "Regression Analysis of Data from Complex Surveys", Roy. Statist. Soc. Conf. Oxford.
- [23] Holt, D., Smith, T.M.F. and Winter, P.O. (1980), "Regression Analysis of Data from Complex Surveys", Jour. Roy. Statist. Soc. A, 143, 474-483.
- [24] Imvrey, P., Sobel, E. and Francis, M. (1980), "Modeling Contingency Tables from Complex Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 213-217.
- [25] Jonrup, H. and Rennermalm, B. (1976), "Regression Analysis in Samples from Finite Population", Scand. Jour. Statist., 3, 33-37.
- [26] Kaplan, B., Francis, I., and Sedransk, J. (1979), "A Comparison of Methods and Programs for Computing Variances of Estimators from Complex Sample Surveys", Proc. Sect. Survey Meth., Amer. Statist. Assoc., 97-100.
- [27] Kish, Leslie and Frankel, M.R. (1970), "Balanced Repeated Replication for Standard Errors", J. Amer. Statist. Assoc., 65, 1071-1094.

- [9] Fienberg, S.E. (1980), "The Measurement of Crime Victimization: Prospects for Panel Analysis of a Panel Survey", The Statistician, 29, 313-350.
- [10] Fuller, W.A. (1975), "Regression Analysis for Sample Survey", Sankhya, 22, 117-132.
- [11] Fuller, W.A. and Battese, C.E. (1973), "Transformations for Estimation of Linear Models with Nested-Error Structure", J. Amer. Statist. Assoc. 68, 626-632.
- [12] Fuller, W.A. and Rao, J.N.K. (1978), "Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix", Ann. Statist. 6, 1149-1158.
- [13] Freeman, D.H. Jr., Freeman, J., Brock, D.B. and Koch, G.G., "Strategies in the Multivariate Analysis of Data from Complex Surveys II: An Application to the United States National Health Interview Survey", Inter. Statist. Rev. 44, 317-330.
- [14] Garza-Hernandez, T. and McCarthy, P.J. (1962), "A Test of Homogeneity for a Stratified Sample", Proc. Soc. Statist. Sect., Amer. Statist. Assoc., 200-202.
- [15] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969), "Analysis for Categorical Data by Linear Models", Biometrics, 25, 489-504.
- [16] Hartley, H.O. and Sielken, R.L. (1975), "A Superpopulation Viewpoint for Finite Population Sampling", Biometrics, 31, 411-422.
- [17] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1980), "Super Camp: Sixth Edition", Statistical Laboratory Survey Section, Iowa State University, Ames, Iowa.

# REMERCIEMENTS

L'auteur tient à remercier MM. D. Binder, N. Chinnappa, S.E. Fienberg, M. Hidiroglou, C.J.C. Hole, J.N.K. Rao et A. Scott de lui avoir consacré du temps et fait part de leurs commentaires.

# BIBLIOGRAPHIE

[1] Altham, P.M.E. (1976), "Discrete Variable Analysis for Individuals Grouped into Families", *Biometrika*, 63, 263-269.

[2] Brewer, K.R. and Mellor, R.W. (1973), "The Effect of Sample Structure on Analytical Surveys", *Aust. J. Statist.*, 15, 145-152.

[3] Bebbington, A.C. and Smith, T.M.F. (1977), "The Effect of Survey Design on Multivariate Analysis", *The Analysis of Survey Data* (C.A. O'Muircheartaigh and C. Payne, Editors), Vol. 2. Model Fitting", New York: Wiley, 175-192.

[4] Campbell, C. (1977), "Properties of Ordinary and Weighted Least Squares Estimators for Two Stage Samples", *Proc. Soc. Statist. Sect., Amer. Statist. Assoc.*, 800-805.

[5] Cohen, J.E. (1976), "The Distribution of the Chi-Squared Statistic Under Clustered Sampling", *J. Amer. Statist. Assoc.*, 71, 665-670.

[6] Cowan, J. and Binder, D.A. (1978), "L'effet d'un plan d'échantillonnage à deux degrés sur les tests d'indépendance", *Techniques d'enquête* 4, n° 1, 16-29.

[7] Fay, R.E. (1979), "On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling", *Proc. Soc. Statist. Sect., Amer. Statist. Assoc.* 402-405.

[8] Fellegi, I.P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", *J. Amer. Statist. Assoc.* 75, 261-268.

### 3.3 Autres genres d'analyses

Les modèles linéaires, de même que les tests de validité de l'ajustement et d'indépendance se prêtent à de nombreuses applications d'analyses, alors que d'autres genres d'analyses telles que l'analyse en composantes principales, l'analyse factorielle et discriminante, l'analyse de corrélation, la régression logistique, les modèles logarithmiques linéaires, les méthodes non paramétriques, etc. ne peuvent être appliquées aussi facilement. Bien que les méthodes générales décrites au chapitre deux puissent être utilisées, leur application pose certaines difficultés et, de fait, on n'a signalé que très peu de cas possibles.

Étant donné que les coefficients de corrélation sont un élément de base dans la plupart des analyses à plusieurs variables, un certain nombre d'études empiriques sur l'effet du plan d'échantillonnage sur l'estimation de ces coefficients ont été menées par Kish et Frankel (1974), Bebbington et Smith (1977) et Holt, Richardson et Mitchell (1980). Quoiqu'on ne puisse tirer de conclusions générales de leurs travaux, il apparaît que les effets du plan d'échantillonnage sont loin d'être négligeables. Bebbington et Smith (1977) ont également étudié la variabilité échantillonnale des estimateurs des composantes principales.

Dans d'autres domaines également, l'effet du plan d'échantillonnage sur les logits a été examiné par Lepkowski et Landis (1980) et les intervalles de confiance des quantiles, par Woodruff (1952) et par Sedransk et Meyer (1978).

où  $p_{ij}$  est la probabilité de distribution de la population dans la case (i, j),  $p_{i+}$ ,  $p_{+j}$  sont les probabilités marginales et  $\bar{p}' = (p_{11}, \dots, p_{rc-1})$ . La statistique généralisée de Wald appliquée au test de  $H_0$  est notée:

$$\chi^2_{MI} = n[\hat{h}(\hat{p})]' \hat{V}^{-1} \hat{h}(\hat{p}), \quad (3.2.8)$$

où  $[\hat{h}(\hat{p})]' = [h_1(\hat{p}), \dots, h_{r-1}(\hat{p})]$  et  $\hat{V}/n$  est un estimateur convergent de la matrice des covariances de  $\hat{h}(\hat{p})$ . Des versions de (3.2.8) ont été appliquées à des plans d'échantillonnage particuliers de même que diverses méthodes d'estimation de  $V_h/n$  ont été utilisées par Garza-Hernandez et McCarthy (1962), Nathan (1969, 1975) Shuster et Downing (1976) et Feillegi (1980).

Une variable modifiée s'apparentant à  $\chi^2/\lambda$  a été introduite par Rao et Scott (1981):

$$\chi^2_{CI} = (n/\hat{\delta}) \sum_{i=1}^r \sum_{j=1}^c (p_{ij} - \hat{p}_{i+} \hat{p}_{+j})^2 / (\hat{p}_{i+} \hat{p}_{+j}), \quad (3.2.9)$$

$$\text{ou } \hat{\delta} = \frac{1}{r} \sum_{i=1}^r \sum_{j=1}^c \hat{V}_{ij}(\hat{h}) / (\hat{p}_{i+} \hat{p}_{+j}) \text{ et}$$

$\hat{V}_{ij}(\hat{h})/n$  est un estimateur convergent de la variance de  $\hat{h}_{ij}(\hat{p})$ .  $\hat{\delta}$  peut être exprimé en fonction des effets du plan estimés de  $\hat{h}_{ij}(\hat{p})$ :

$$\hat{\delta} = \frac{1}{r} \sum_{i=1}^r \sum_{j=1}^c (1 - \hat{p}_{i+}) (1 - \hat{p}_{+j}) \hat{\delta}_{ij}, \quad (3.2.10)$$

où  $\hat{\delta}_{ij}$  est un estimateur de l'effet du plan d'échantillonnage, de  $\hat{h}_{ij}(\hat{p})$ :

$$\hat{\delta}_{ij} = nV[h_{ij}(\hat{p})] / [p_{i+} p_{+j} (1 - p_{i+}) (1 - p_{+j})]. \quad (3.2.11)$$

Il peut être plus facile d'estimer les effets du plan d'échantillonnage que d'estimer les variances.

Des travaux de recherche empiriques menés par Holt, Scott et Ewings (1980) et par Hidiroglou et Rao (1981) démontrent que la distribution de  $\chi^2_{CI}$  s'apparente à  $\chi^2_{(r-1)(c-1)}$ .



où  $\lambda_1, \dots, \lambda_{k-1}$  sont les valeurs propres de

$$D = P_0^{-1} V (\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{k-1} > 0).$$

(3.2.6)

On peut alors procéder à un test prudent de (3.2.1) en utilisant la variable  $\chi^2/\lambda_1$  et une distribution  $\chi^2_{k-1}$ .  $\lambda_1$  peut représenter l'effet maximum du plan sur toutes les combinaisons linéaires des composantes de  $\bar{p}$ . Par exemple, avec un échantillonnage stratifié proportionnel,  $\lambda_1 \leq 1$  de sorte que  $\chi^2$  peut être utilisé comme une fonction conservative des observations dans le test.

Dans les autres cas, l'utilisation de  $\chi^2/\lambda$

$$\text{soit } \lambda = \frac{1}{k-1} \sum_{i=1}^{k-1} \lambda_i = \frac{1}{k-1} \sum_{i=1}^k d_i (1-p_i),$$

où  $d_i = V[p_i]/[p_i(1-p_i)]$  est l'effet du plan pour  $p_i$ , s'est révélée un bon test d'approximation (Hidiroglou et Rao (1981), pour l'Enquête Santé Canada, et Holt, Scott et Ewings (1980), pour des enquêtes à grande échelle au

Royaume-Uni). Une

approximation alternative,  $\chi^2/\bar{d}$ , où  $\bar{d} = k^{-1} \sum_{i=1}^k d_i$ , a été

proposée par Felllegi (1980).

La formulation directe de modèles pour  $\bar{p}$  a été présentée par Altham (1976) et par Cohen (1976), mais leurs modèles comportent de sérieuses limitations puisqu'ils supposent que  $\lambda_1 = \lambda_2 = \dots = \lambda_{k-1} = \bar{\lambda}$ , ce qui équivaut à un effet constant du plan sur les catégories. Cela ne constitue pas généralement une hypothèse réaliste et a pour résultat que  $\chi^2/\bar{\lambda}$  contient une distribution asymptotique  $\chi^2_{k-1}$ . Pour effectuer un test d'indépendance dans une table de contingence, les hypothèses peuvent être exprimées:

$$H_0: h_{ij}(p) = p_{ij} - p_{i+} p_{+j} = 0$$

$$(i=1, \dots, r-1; j=1, \dots, c-1),$$

(3.2.7)

$\bar{p}' = (p_1, \dots, p_{k-1})$ . Lorsqu'on veut tester l'hypothèse nulle de la validité de l'ajustement d'une distribution connue,  $\bar{p}_0 = (p_{01}, \dots, p_{0k-1})$ :

$$(3.2.1) \quad H_0: \bar{p} = \bar{p}_0,$$

on peut suivre les démarches expliquées au chapitre 2.

Nous supposons qu'il existe un estimateur d'enquête convergent  $\hat{\bar{p}}' = (\hat{p}_1, \dots, \hat{p}_{k-1})$  de  $\bar{p}'$ . S'il est asymptotiquement normal:

$$(3.2.2) \quad \sqrt{n}(\hat{\bar{p}} - \bar{p}) \sim N(0, V)$$

et qu'il existe un estimateur convergent  $\hat{V}$  de  $V$ , alors la statistique généralisée de Wald:

$$(3.2.3) \quad \chi_W^2 = n(\hat{\bar{p}} - \bar{p}_0)' \hat{V}^{-1}(\hat{\bar{p}} - \bar{p}_0),$$

qui est distribuée asymptotiquement par  $\chi_{k-1}^2$  sous  $H_0$ , peut être utilisée pour tester  $H_0$ .

Pour bon nombre de plans d'échantillonnage simples, des estimateurs convergents de  $V$  sont obtenus directement alors que, pour les plans plus complexes, ils peuvent être calculés à l'aide des méthodes classiques. Cependant, si des tests de validité de l'ajustement doivent être effectués pour une variété de variables et de classifications, il est alors préférable d'utiliser la variable  $\chi^2$  type:

$$(3.2.4) \quad \chi^2 = n \sum_{i=1}^k (p_i - p_{0i})^2 / p_{0i} = n(\hat{\bar{p}} - \bar{p}_0)' \hat{P}_0^{-1}(\hat{\bar{p}} - \bar{p}_0),$$

où  $\bar{P}_0 = \text{diag}(\bar{p}_0) - \bar{p}_0 \bar{p}_0'$ , tout en y apportant les modifications qui s'imposent. Rao et Scott (1981) ont démontré que la distribution asymptotique  $\chi^2$  sous  $H_0$  est celle de la somme pondérée de  $k-1$  variables indépendantes  $X$  ayant chacune un degré de liberté.

$$(3.2.5) \quad \chi^2 \sim \sum_{i=1}^{k-1} \lambda_i Z_i^2; \quad Z_i \sim N(0,1) \text{ indépendantes}$$

Le programme de régression pondérée, dont les coefficients sont  $1/\Pi_1$ , définit une valeur de  $(X_n' W_n X_n)^{-1}$  de la variance du modèle de  $\hat{\beta}_W$ , ce qui correspond à (3.1.11) seulement si  $W_n = I_n$ . Par conséquent, aucun résultat concernant les erreurs types ou les tests d'hypothèses n'est exact.

Toutefois, l'estimateur du coefficient de corrélation multiple défini à l'aide de la régression pondérée:

$$\hat{R}_2^2 = \frac{(Y_n - \hat{X}_n \hat{\beta}_W)' W_n (Y_n - \hat{Y}_n)}{(Y_n - \hat{Y}_n)' W_n (Y_n - \hat{Y}_n)} \quad (3.1.12)$$

où  $\hat{Y}_n = (Z_s' 1/\Pi_1) / (Z_s' 1/\Pi_1)$  est un estimateur convergent basé sur un plan d'échantillonnage du coefficient de corrélation multiple de la population:

$$R_2^2 = \frac{(Y - X B)' (Y - \hat{Y})}{(Y - X B)' (Y - \hat{Y})} \quad (3.1.13)$$

$$\text{où } \hat{Y} = (1/N) 1' Y \cdot$$

La variance basée sur un plan de  $\hat{\beta}_W$ , lequel doit être considéré comme la mesure appropriée de la précision de  $\hat{\beta}_W$  en tant qu'estimateur de  $B$ , ne peut pas en général être calculée uniquement à l'aide des probabilités d'inclusion de premier ordre  $\Pi_1$ . Pour la plupart des plans

d'échantillonnage appliqués couramment, la variance de  $\hat{\beta}_W$  basée sur un plan est estimée à l'aide d'une des méthodes d'estimation de la variance mentionnées plus haut, soit la linéarisation, la répétition équilibrée ou le "jackknife" (voir Jonrup et Renneralm (1976) et Holt et Scott (1981)).

### 3.2 Analyse des données qualitatives

La méthode la plus simple d'analyse des données qualitatives est celle qui consiste à construire une classification unique de la population en  $k$  classes comportant des probabilités (fréquences relatives)

De toute façon,  $\hat{\beta}_W$  est un estimateur sans biais de  $\beta$  basé sur le modèle, chaque fois que le modèle (3.1.4) se vérifie. Règle générale, il ne sera pas un estimateur optimal de  $\beta$  selon (3.1.5) lorsqu'il y a échantillonnage à probabilité inégale, mais il le sera si la variance conditionnelle du modèle de  $Y_i$  est proportionnelle à  $\Pi_i$ , soit:

$$V(Y_i | x_i) = k \Pi_i \quad (3.1.9)$$

Comme l'estimateur pondéré  $\hat{\beta}_W$  est plus robuste que l'estimateur non pondéré  $\hat{\beta}$ , en ce sens qu'il est à la fois un estimateur sans biais de  $\beta$  basé sur le modèle, si le modèle est vrai, et un estimateur convergent de  $\beta$  basé sur le plan d'échantillonnage, si le modèle ne se vérifie pas, il est recommandé d'utiliser l'estimateur pondéré  $\hat{\beta}_W$  pour estimer  $\beta$ , chaque fois qu'on n'est pas sûr si le modèle (3.1.4) - (3.1.5) se vérifie. Il reste alors aux spécialistes à déterminer si  $\beta$  est un paramètre valable à estimer.

Il convient de souligner que, dans le cas de plans d'échantillonnage autopondérés,  $\hat{\beta}$  et  $\hat{\beta}_W$  correspondent. L'estimateur  $\hat{\beta}_W$  (3.1.8) peut être calculé directement à partir de programmes informatiques types qui définissent la régression pondérée (par exemple BMDP) à l'aide des coefficients de pondération  $1/\Pi_i$ , ou encore à l'aide d'autres programmes (par exemple SPSS) qui affectent la régression non pondérée aux variables transformées  $Y_i/\sqrt{\Pi_i}$  et  $x_i/\sqrt{\Pi_i}$ , mais pas aux variables pondérées  $Y_i/\Pi_i$ ,  $x_i/\Pi_i$ . Cependant, il est bon de noter que, sous l'hypothèse alternative, les variances et covariances des estimateurs sont incorrectes et que les tests de signification usuels (par exemple, les tests  $F$ ) ne sont pas valides et peuvent fausser grandement les conclusions.

Soit le modèle (3.1.4) - (3.1.5), la variance du modèle de  $\hat{\beta}$  est:

$$V(\hat{\beta} | X) = \sigma^2 (X' X)^{-1} \quad (3.1.10)$$

ce qui est le résultat fourni par les programmes de régression non pondérée. Cependant, la variance du modèle  $\hat{\beta}_W$  est:

$$V(\hat{\beta}_W | X) = \sigma^2 (X' W X)^{-1} \quad (3.1.11)$$

basée sur les valeurs échantillonnées

$$X'_n = (x_1, \dots, x_n) \text{ et } Y'_n = (y_1, \dots, y_n)$$

produit le "meilleur" estimateur de  $\beta$  sans biais d'un modèle linéaire, peu importe le plan d'échantillonnage. On le qualifie de meilleur parce qu'il définit une variance minimum basée sur le modèle. Toutefois,  $\beta$  n'est pas, de façon générale, un estimateur sans biais basé sur le plan d'échantillonnage, ni même un estimateur convergent (basé sur le plan) du paramètre de la population:

$$B = (X'_N X_N)^{-1} X'_N Y_N$$

(3.1.7)

$$\text{ou } X'_N = (x_1, \dots, x_N) \text{ et } Y'_N = (y_1, \dots, y_N)$$

L'estimateur convergent de  $B$  selon le plan d'échantillonnage est l'estimateur pondéré:

$$\hat{\beta}_W = (X'_W W X_W)^{-1} X'_W W Y_W$$

(3.1.8)

où la matrice de pondération  $W_n = \text{diag}(\pi_1, \dots, \pi_n)$ , est la matrice diagonale  $n \times n$  des inverses des probabilités d'inclusion de l'échantillon  $\pi_i = p_i$  (ies)

De toute évidence, la compatibilité de  $\hat{\beta}_W$ , comme estimateur de  $\beta$ , ne dépend pas de ce que le modèle (3.1.4) se vérifie; en outre, la pertinence de l'estimation de  $B$ , lorsque le modèle ne se vérifie pas, peut être mise en doute. On peut démontrer que, si certaines conditions se réalisent dans un modèle non-linéaire, ce qui suppose que l'espérance mathématique conditionnelle de  $Y$  (selon  $X$ ) est une fonction différentiable de  $X$ , l'espérance de  $B$  basée sur le modèle peut être exprimée approximativement comme une moyenne pondérée des pentes de cette fonction aux points  $X_i$  (les coefficients de pondération dépendant seulement de  $X_i - \bar{X}$ ). Cependant, dans la pratique, cette interprétation a une valeur limitée.



plan d'échantillonnage complexe, c'est-à-dire que la distribution de l'échantillonnage dépend seulement de  $X_2$ :

$$(3.1.2) \quad P(s|X_1, X_2) = P(s|X_2).$$

L'estimation de  $\beta_{1.2}$  et de  $\beta_{2.1}$  dans la définition (3.1.1) et l'inférence qui en est établie peuvent être calculées selon la méthode classique, comme si l'échantillonnage était aléatoire simple, pour autant que le terme (3.1.1) se vérifie.

Cependant, si les variables du plan,  $X_2$ , ne sont pas incluses dans l'équation de régression d'intérêt:

$$(3.1.3) \quad E(Y|X_1) = X_1\beta_1$$

et que la variable du plan,  $X_2$ , est en corrélation avec  $Y$  (fonction conditionnelle de  $X_1$ ), alors l'estimation type par la méthode des moindres carrés ordinaires de  $\beta_1$  n'est pas convergente (voir Nathan et Holt (1980) et Holt et Smith (1979) qui préconisent l'utilisation d'estimations modifiées pondérées et non pondérées de  $\beta_1$ , qui sont convergentes). Holt, Smith et Winter (1980) donnent un exemple d'application de ces estimateurs.

Si le modèle linéaire:

$$(3.1.4) \quad E(Y_i|X_i) = X_i'\beta$$

$$(3.1.5) \quad \text{cov}(Y_i, Y_j | X_i, X_j) = \begin{cases} \sigma^2 & i=j \\ 0 & i \neq j \end{cases}$$

est vraie pour toutes les unités ( $i, j=1, \dots, N$ ) d'une population finie et  $X_i'$ , le vecteur colonne  $p \times 1$ , comprend toutes les variables du plan d'échantillonnage, alors l'estimation non pondérée des moindres carrés ordinaires:

$$(3.1.6) \quad \hat{\beta} = (X'X)^{-1}X'Y$$

La rectification peut être basée sur les effets du plan de divers estimateurs ou sur la moyenne des effets du plan (voir Cowan et Binder (1978), Fay (1979), Fallægj (1980), Rao et Scott (1981) et Holt (1981).

L'autre solution possible est d'étudier le comportement des fonctions des observations utilisées dans un test selon un certain modèle de superpopulation et de modifier en conséquence les fonctions types (Cohen, 1979, et Campbell, 1977).

### 3. METHODES PARTICULIÈRES

#### 3.1 Régression et modèles linéaires

La définition préalable du modèle et des paramètres d'intérêt est

extrêmement importante dans le cas de l'analyse de la régression et des

modèles linéaires. Ainsi, lorsque des relations de régression différentes

doivent être calculées pour diverses strates ou UPE dans un plan

d'échantillonnage à deux degrés, le paramètre d'intérêt peut être une simple

moyenne des coefficients de régression (Konijn, 1962), une moyenne pondérée

des coefficients (Pfeffermann et Nathan, 1981) ou leur valeur prévue (selon

une certaine distribution préalable - Porter, 1973).

Règle générale, le modèle et les paramètres d'intérêt doivent être définis

en fonction de la structure présupposée de la population globale et ne doivent

pas se rapporter à la structure du plan d'échantillonnage. Toutefois, dans

de nombreux cas, le plan d'échantillonnage reproduit la structure de la

population de sorte que les variables du plan font partie du modèle. A

titre d'exemple, considérons le modèle:

$$E(Y|X_1, X_2) = X_1 \beta_{1.2} + X_2 \beta_{2.1} \quad (3.1.1)$$

où  $X_1$  comprend uniquement les variables qui ne se rapportent pas au plan d'échantillonnage et  $X_2$  représente toutes les variables incluses dans le

## 2.2 Approximation et élaboration d'un modèle de covariances

Les problèmes courants liés au calcul d'un estimateur convergent stable de la matrice des covariances ont poussé les spécialistes à essayer d'appliquer des approximations simplifiées à ces estimateurs. Le principe de base est que si l'on suppose une certaine structure de la matrice des covariances, on peut alors utiliser des estimateurs plus stables d'un petit nombre de paramètres.

L'approximation peut être calculée à l'aide de la méthode basée sur un plan d'échantillonnage, directement en fonction de la matrice des covariances. Si l'on peut faire des hypothèses sur l'égalité des effets du plan d'échantillonnage sur les variances et covariances à l'intérieur d'un sous-groupe donné de paramètres, les estimateurs d'ensemble de la covariance peuvent alors être utilisés. Cette démarche est préconisée par Nathan (1973), Rao (1978), Fellegi (1980) et Lepkowski et Landis (1980).

Par ailleurs, l'élaboration d'un modèle de la structure de la population proprement dite peut favoriser la construction de matrices de covariances simplifiées qui peuvent être facilement estimées (voir Altham (1976), Fuller et Battese (1973), Tomberlin (1979), Holt, Richardson et Mitchell (1980), Imrey, Sobel et Francis (1980) et Pfeffermann et Nathan (1981).

## 2.3 Modification des tests classiques

L'utilisation très répandue de programmes informatiques types a favorisé l'élaboration de nouveaux tests qui tiennent compte de plans d'échantillonnage complexes. Ces travaux peuvent être évalués sous l'angle d'une extension naturelle de l'utilisation des effets du plan d'échantillonnage comme facteurs multiplicatifs des variances basées sur un échantillon aléatoire simple d'une même taille, en vue d'apporter les modifications appropriées lorsqu'il s'agit de plans d'échantillonnage complexes.

est distribuée asymptotiquement, selon l'hypothèse nulle, comme  $\chi^2$  ayant des degrés de liberté correspondant à la dimension de l'hypothèse (p-r).

La compatibilité de  $\hat{\beta}$  et de  $\hat{V}$  et les distributions asymptotiques de  $\hat{\beta}$  et de  $\chi^2_W$  peuvent être étudiées en fonction de la distribution de la

superpopulation.

La principale difficulté que présente cette démarche a trait au calcul de l'estimateur convergent  $V$  de la matrice des covariances, lorsque  $\beta$  n'est pas linéaire dans les données de l'échantillon (ce cas est très fréquent). Rao (1975) a étudié les diverses méthodes d'estimation de la variance qui peuvent être appliquées, notamment la linéarisation (Teppling, 1968), la répétition équilibrée (McCarthy, 1969) et la méthode dite "Jackknife" (Miller, 1974). Il existe également plusieurs programmes informatiques qui peuvent être utilisés dans ces cas, par exemple SUPERCARP (Hidiroglou, Fuller et Hickman, 1980), SUDAAN (Shah, 1978) pour la linéarisation et OSIRIS IV: PSALMS pour la répétition équilibrée. Une liste complète et une comparaison des programmes est présentée par Kaplan, Francis et Sedransk (1979).

Des comparaisons empiriques des estimateurs de la variance ont été établies par Kish et Frankel (1974) et par Richards et Freeman (1980), de même que des comparaisons théoriques, par Krewski et Rao (1981).

Toutefois, il convient de porter une attention particulière à la stabilité de l'estimateur de la variance, surtout lorsqu'il y a un grand nombre de paramètres. De plus, il faut aussi tenir compte des conditions dans lesquelles la convergence et les propriétés asymptotiques se réalisent, dans le cas des plans complexes. Par exemple, avec un plan d'échantillonnage à deux degrés, il est possible que les résultats asymptotiques commandent un grand nombre d'unités primaires d'échantillonnage (UPC) et un grand nombre d'unités du dernier degré par UPC.



d'appliquer soit la méthode basée sur un plan d'échantillonnage soit celle basée sur un modèle, selon le degré de confiance que chacun a dans la validité d'un modèle sous-jacent.

## 2. METHODES GÉNÉRALES DE BASE

### 2.1 Statistique généralisée de Wald

Si l'hypothèse qu'il faut tester est linéaire (ou peut être linéarisée) dans les valeurs prévues de statistiques asymptotiquement normales, pour lesquelles il existe un estimateur convergent de la matrice des variances, on peut alors utiliser la statistique généralisée de Wald (Grizzle, Starmer et Koch (1969), Koch, Freeman et Freeman (1976), Freeman, Brock et Koch (1976), Shah, Holt et Folsom (1977) et Koch, Stokes et Brock (1980)).

Nous supposons que nous voulons tester l'hypothèse:

$$H_0: X\beta = \theta_0, \quad (2.1.1)$$

où  $X$  est une matrice  $r \times p$  connue de plein rang du plan d'échantillonnage.  $\beta$  est un vecteur inconnu  $p \times 1$  des paramètres (soit des paramètres d'une population finie soit des paramètres d'une superpopulation) et  $\theta_0$  est un vecteur connu  $r \times 1$  des constantes. Si l'hypothèse n'est pas linéaire, on peut procéder à une approximation de premier ordre des séries de Taylor (Nathan, 1972, et Shuster et Downing, 1976).

Nous supposons qu'il y a un estimateur asymptotique normal convergent  $\hat{\beta}$  de  $\beta$  ainsi qu'un estimateur convergent ( $V$ ) de la matrice des covariances de  $\hat{\beta}$ , dont la distribution est indépendante de celle de  $\hat{\beta}$ .

Alors, la statistique généralisée de Wald définie par:

$$\chi^2_W = (\hat{X}\hat{\beta} - \theta_0)' (X'VX)^{-1} (\hat{X}\hat{\beta} - \theta_0) \quad (2.1.2)$$



qu'un grand nombre de petits résidus présentant des coefficients de pondération élevés. Un des outils d'analyse particulièrement utile dans le cas de la régression est la différence entre les coefficients de régression pondérée et non pondérée. Un écart important démontre souvent que le modèle ne convient pas.

Une fois les paramètres définis, il faut déterminer le genre d'inférence, soit l'estimation ponctuelle, l'estimation par intervalle ou les tests d'hypothèses. Il apparaît que l'estimation ponctuelle et les intervalles de confiance conviennent très bien aux paramètres d'une population finie, alors que les tests d'hypothèses, plus particulièrement d'hypothèses simples, sont valables uniquement lorsqu'ils sont appliqués à des paramètres d'une superpopulation d'un modèle très bien défini. Par exemple, l'hypothèse que les moyennes d'une paire d'ensembles sont égales peut être admise seulement s'il s'agit des moyennes d'une superpopulation et non des réalisations de leur population finie. Lorsqu'on veut éviter la formulation d'un modèle, il est recommandé de procéder à l'estimation ponctuelle ou de calculer les intervalles de confiance, plutôt que de faire des tests d'hypothèses, pour établir la différence entre les moyennes d'un ensemble. S'il faut appliquer ces tests aux paramètres d'une population finie, il est préférable de tester une hypothèse composée (par exemple tester si la différence entre les moyennes se situe à l'intérieur d'un ordre de valeurs) que de tester une hypothèse simple, c'est-à-dire vérifier si la différence est zéro. Il convient de souligner que, dans le cas d'échantillons assez grands, tout écart plus grand que zéro, peu importe s'il l'est à peine, sera considéré comme différent de zéro de façon significative.

Nous allons maintenant examiner certaines méthodes générales d'analyse des données provenant de plans d'échantillonnage complexes, de même que certaines autres méthodes particulières concernant les modèles linéaires et les tests de validité de l'ajustement et d'indépendance dans les tables de contingence. De façon générale, nous étudierons l'inférence se rapportant aux paramètres d'une population finie. Cependant, nous estimons que cette inférence est pertinente seulement si les paramètres s'apparentent aux paramètres du modèle de superpopulation, ce qui laisse à chacun la liberté

Ces deux démarches peuvent soulever de vives objections, car la méthode basée sur un modèle repose en grande partie sur des hypothèses appliquées à un modèle théorique qui ne peuvent généralement pas être garanties; par voie de conséquence, l'inférence ne sera pas robuste s'il y a déviation de ce modèle. Par ailleurs, dans le cas de l'inférence basée sur un plan d'échantillonnage, les paramètres d'une population finie sont habituellement des "copies" des paramètres d'un modèle théorique qui présentent une valeur descriptive très faible, à moins qu'il ne s'agisse d'un modèle de base. Par exemple, le coefficient de corrélation d'une population finie peut être une mesure utile de la relation qui existe entre deux variables, mais seulement si cette relation est approximativement linéaire.

Dans de nombreux cas, il est préférable de chercher un certain équilibre entre ces deux méthodes d'inférence. Il est possible d'y arriver si l'on utilise uniquement les paramètres d'une population finie qui constituent des éléments approximatifs des paramètres de la superpopulation d'un modèle pertinent auquel les données peuvent être appliquées. Par exemple, lorsque des équations de régression distinctes sont ajustées aux sous-ensembles appropriés, on obtient alors un meilleur ajustement linéaire que si l'on procède à une régression générale. Si les sous-ensembles sont de taille assez grande, les coefficients de régression de la population finie pouront mieux s'apparenter aux paramètres de la superpopulation et, ainsi, l'inférence portant sur les paramètres d'une population finie pourra être considérée comme caractérisant les paramètres de la superpopulation.

Pour pouvoir assurer la meilleure correspondance possible entre les paramètres d'un modèle et ceux d'une population finie, il faut procéder à une analyse exploratoire en profondeur du modèle avant d'aborder l'analyse théorique. Cette analyse exploratoire des divers modèles alternatifs peut, dans la plupart des cas, être fondée sur des mesures descriptives pour lesquelles le plan d'échantillonnage peut être pris en compte, ou encore sur des représentations graphiques. Toutefois, les résultats doivent être interprétés avec prudence en fonction du plan d'échantillonnage. Par exemple, un petit nombre de grands résidus ayant des coefficients de pondération de l'échantillon faibles peut être beaucoup moins important

Dans le présent article, nous tentons de donner quelques conseils pratiques sur ce qui peut et ne doit pas être fait dans ces cas. Cet examen est fondé sur des travaux récents (voir bibliographie) portant sur ces questions et comprend plusieurs exemples d'applications possibles.

Quiconque veut entreprendre une analyse statistique doit tout d'abord définir les paramètres qu'il faut estimer. Pour ce faire, on peut s'inspirer des travaux de Brewer et Mellor (1973) et de Smith (1976). De leur côté, Kish et Frankel (1974) soutiennent qu'une inférence valable doit porter sur des paramètres de la population finie, notamment le coefficient de régression d'une population :

$$B = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2},$$

ou les coefficients de corrélation multiple ou partielle ou d'autres mesures qui sont définies par rapport à la population finie seulement, sans utiliser un modèle de superpopulation. Dans ce cas, l'inférence est basée sur un plan d'échantillonnage (Särndal, 1978), c'est-à-dire qu'elle est fondée uniquement sur les propriétés de la distribution d'échantillonnage. Cependant, on peut aussi utiliser l'inférence basée sur un modèle pour estimer un paramètre de la population finie (Hartley et Sielken, 1975).

Il existe une autre possibilité d'estimation dans ce domaine, celle que propose par exemple Fienberg (1980). Selon lui, l'inférence doit porter sur les paramètres d'une distribution de probabilité (superpopulation) dont la population finie constitue la réalisation. On peut trouver des exemples de ce type d'inférence dans les travaux de Konijn (1962), Fuller (1975), Thomsen (1978) et Pfeffermann et Nathan (1981). Si les paramètres se rapportent à un modèle de superpopulation, l'inférence ne peut être basée uniquement sur un plan d'échantillonnage; elle doit être fondée sur un modèle (Särndal, 1978) ou encore, sur un modèle et un plan d'échantillonnage en même temps. Si l'on suppose qu'il y a indépendance entre la distribution du modèle et la distribution d'échantillonnage, l'inférence classique (basée sur un modèle) est alors valide et le plan d'échantillonnage peut seul avoir un effet sur l'efficacité de l'inférence.



# L'INFÉRENCE STATISTIQUE BASÉE SUR DES PLANS D'ÉCHANTILLONNAGE COMPLEXES

Gad Nathan<sup>1</sup>

Les problèmes d'inférence statistique basée sur des plans d'échantillonnage complexes sont décrits. La définition du paramètre d'intérêt est de grande importance et on doit décider si on veut estimer un paramètre de la population à nombre fini ou un paramètre du modèle de superpopulation. Les méthodes générales basées sur la statistique généralisée de Wald et sa modification, aussi bien que la modification des statistiques classiques sont présentées. Les méthodes spécifiques pour la régression et les modèles linéaires et pour l'analyse des données qualitatives sont décrites en détail.

## 1. INTRODUCTION

Règle générale, l'application des méthodes classiques d'inférence, telles que la régression, l'analyse de variance ou les tests d'hypothèses, suppose que les observations proviennent d'un échantillon aléatoire simple d'une population infinie ayant une distribution de probabilité particulière à une famille hypothétique. La grande diffusion des programmes informatiques a rendu l'utilisation de ces méthodes particulièrement facile. Toutefois, il est presque impossible de les appliquer telles quelles à des données recueillies au moyen de plans d'échantillonnage complexes.

<sup>1</sup> G. Nathan, Hebrew University, Jérusalem et Bureau de la statistique d'Israël.





Préparé par les méthodologistes de Statistique Canada

## Comité de rédaction:

R. Platek - Président  
M.P. Singh - Rédacteur en chef  
P.F. Timmons

J.H. Gough - Rédacteur adjoint

## Politique de la rédaction:

La revue "Techniques d'enquête" veut donner aux personnes qui s'intéressent aux aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquête: les problèmes de conception causés par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada.

## Présentation de documents pour publication

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement, et d'enquêtes ménages Statistique Canada, 6<sup>e</sup> étage, Edifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi.



Préparé par les méthodologistes de Statistique Canada.

TABLE DES MATIÈRES

109	L'inférence statistique basée sur des plans d'échantillonnage complexes GAD NATHAN .....
-----	--

131	Le problème de la non-réponse J.G. BETHLEHEM et H.M.P. KERSTEN .....
-----	---

162	Les variances d'estimateurs asymptotiquement normaux basés sur des enquêtes complexes DAVID A. BINDER .....
-----	---

179	La statistique de la santé au Canada: rétrospective et jalons pour l'avenir LORNE ROWEBOTTOM .....
-----	--

187	Modèle les d'estimation des erreurs d'échantillonnage P.D. GHANGURDE .....
-----	---



# TECHNIQUES D'ENQUÊTE

décembre 1981

volume 7

numéro 2

Préparé par les  
méthodologistes de  
Statistique Canada

Canada





2-001



Statistics Canada Statistique Canada

Government  
Publications

---

# **SURVEY METHODOLOGY**

---

---

**1982**

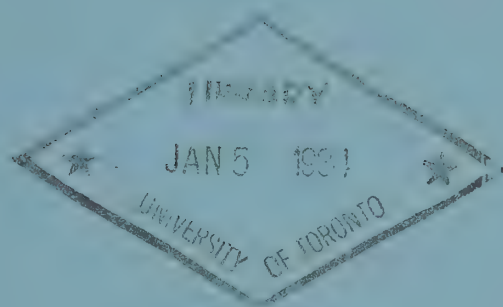
---

**Volume 8**

---

**Numbers 1 & 2**

---



---

A Journal produced by  
Methodology Staff  
Statistics Canada

---

Canada



SURVEY METHODOLOGY

1982

Vol. 8

Nos. 1 & 2

A Journal produced by Methodology Staff, Statistics Canada

CONTENTS

The Role of the Questionnaire in Survey Design R. PLATEK and D. ROYCE .....	1
Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey J.D. DREW, M.P. SINGH and G.H. CHOUDHRY .....	17
Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey ELIZABETH CLAYTON PAUL and MURRAY LAWES .....	48
Rotation Group Bias in the LFS Estimates P.D. GHANGURDE .....	86
Computerization of Complex Survey Estimates M.A. HIDIROGLOU .....	102

8-3200-501  
Reference No.  
Z - 079

ISSN: 0714-0045



## SURVEY METHODOLOGY

1982

Vol. 8

Nos. 1 & 2

A Journal produced by Methodology Staff, Statistics Canada.

---

Editorial Board:	R. Platek	- Chairman
	M.P. Singh	- Editor
	P.F. Timmons	
	J.H. Gough	- Assistant Editor

---

### Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department.

### Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.





## THE ROLE OF THE QUESTIONNAIRE IN SURVEY DESIGN

R. Platek and D. Royce<sup>1</sup>

The modern statistical survey is an effective method of meeting the ever-increasing demand for timely and accurate data. One important component of the statistical survey is the questionnaire. This article discusses the role of the questionnaire in meeting the needs of users, the relationship of the questionnaire to the other components of survey design, and the effect of the questionnaire on the quality of survey data. The importance of viewing the questionnaire as an integral part of the total survey design is stressed.

### 1. INTRODUCTION

The escalating demand for appropriate and timely information of various kinds and from various sources calls for an organized approach to the entire process of data collection. The past forty years have seen the emergence of the statistical survey as an important tool to meet this need.

One important component of the statistical survey is the questionnaire. In the sections which follow, we describe the role of the questionnaire in meeting information needs, the relationship of the questionnaire to the other components of survey design, and the effect of the questionnaire on the quality of survey data. Although the discussion is presented mainly in the context of the household survey conducted by personal interview, many of the comments are relevant to questionnaires and surveys of all types.

### 2. INFORMATION NEEDS AND THE ROLE OF THE QUESTIONNAIRE

The simplest definition of a questionnaire is that of a group or sequence of questions designed to elicit information upon a subject from a respondent. Within the range of techniques in questioning, the questionnaire may range from a list of undefined topics to a highly structured set of questions with no options for response other than those listed.

<sup>1</sup>R. Platek and D. Royce, Census and Household Survey Methods Division Statistics Canada.

The questionnaire plays a central role in a complex process (the interview) in which information is transferred from those who have it (the respondents) to those who need it (the users). The questionnaire is the means through which the information needs of the users are expressed in operational terms which can be presented to a respondent in such a way that he will supply the required information. For this transfer of information to be effective, the questionnaire must meet the requirements of both users and respondents.

The expression of information needs, which a user may initially only vaguely understand, in terms suitable to the respondent is not something that can be accomplished in one step. Instead, the questionnaire design evolves and is refined as part of the overall survey development process.

For example, the user may begin with a need for information on "the housing conditions of the poor". He develops this into survey objectives by asking questions such as:

- (a) What is the problem we are trying to solve?
- (b) What specific items of information are needed?
- (c) How will the information be used?
- (d) How accurate and timely does the information have to be?

In answering these questions, his thinking becomes more quantitative, and he expresses his information needs in terms of specific survey concepts. The survey concepts describe both what is to be measured and the units for which measurements are required. He may describe "housing conditions" in terms of the number of rooms, the presence of plumbing and electricity, or the state of repair of the dwelling. He may define "the poor" in terms of income level or in terms of assets and debts.

It is important to emphasize that specific question wording is not at issue in the development of survey concepts. The first step for the user in expressing

his information needs is to decide what should be measured, not how it will be measured. The user should choose the concepts based on their relevance to his information needs. He should consider, for example, what concepts are most appropriate for the uses to be made of the data and whether the concepts are compatible with other sources of information.

Once information needs have been expressed in terms of specific survey concepts, the questionnaire becomes the instrument by which these concepts are measured. Through specific questions and accompanying instructions, the user specifies precisely how the survey concepts are to be measured in operational terms. Several questions may be required to measure complex concepts. In the Canadian Labour Force Survey, for example, as many as ten questions are needed to measure the concept "unemployed".

The questionnaire often serves as the document for recording of measurements as well. This is mainly of benefit to the interviewer or respondent, since it is convenient to record the answers immediately following the question. In theory, however, there is no reason why the questions and answers cannot be on two separate forms.

In the more structured types of surveys, the questionnaire is an important method of standardizing and controlling the data collection process. In statistical surveys, in contrast to other methods of investigation, the researcher usually cannot do his own data collection but must rely on interviewers hired for the job. Without specific question wordings and instructions to follow, interviewers would inevitably change the meaning or emphasis of questions and quite possibly the responses. The questionnaire helps ensure that the researcher measures what he wishes to measure with every respondent. It is, in effect, a "program" for the interviewer and respondent to follow in order to produce the desired result.

The questionnaire cannot be too rigid, however. It must be flexible enough to adapt to respondents of different age/sex groups, languages and social backgrounds. Different words or groups of words may be needed in order to convey the desired meaning to all respondents. The questionnaire must also

anticipate all of the possible answers that could be given. This is especially true in the initial, exploratory stages of research where an unstructured collection of data may be the most appropriate approach.

It must be recognized that the questionnaire is a complex and often imprecise measuring instrument. The subjects of measurement are human beings, and the process of measurement is based on language. As well as being a measuring instrument, the questionnaire is also a form of communication involving the researcher, the interviewer and the respondent. It transmits a request for information to the respondent, and it transmits the respondent's answer back to the researcher in a form useable to him. Warren Weaver, in The Mathematical Theory of Communication (1949), identifies three problems that must be faced in the design of any communication system:

- A. How accurately can the symbols of communications be transmitted? (The technical problem).
- B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem).
- C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

All three problems are directly relevant to the construction of questionnaires, and all three problems are closely linked. Within the context of statistical surveys, the way in which the questionnaire solves these problems plays a major role in determining how well the information needs of the user are met.

### 3. THE QUESTIONNAIRE AND THE COMPONENTS OF SURVEY DESIGN

The process of making the survey concepts operational in a specific document forces the researcher to consider not only question wording, sequencing and layout, but nearly every other aspect of the survey as well. The questionnaire design must take into account elements such as the type of population



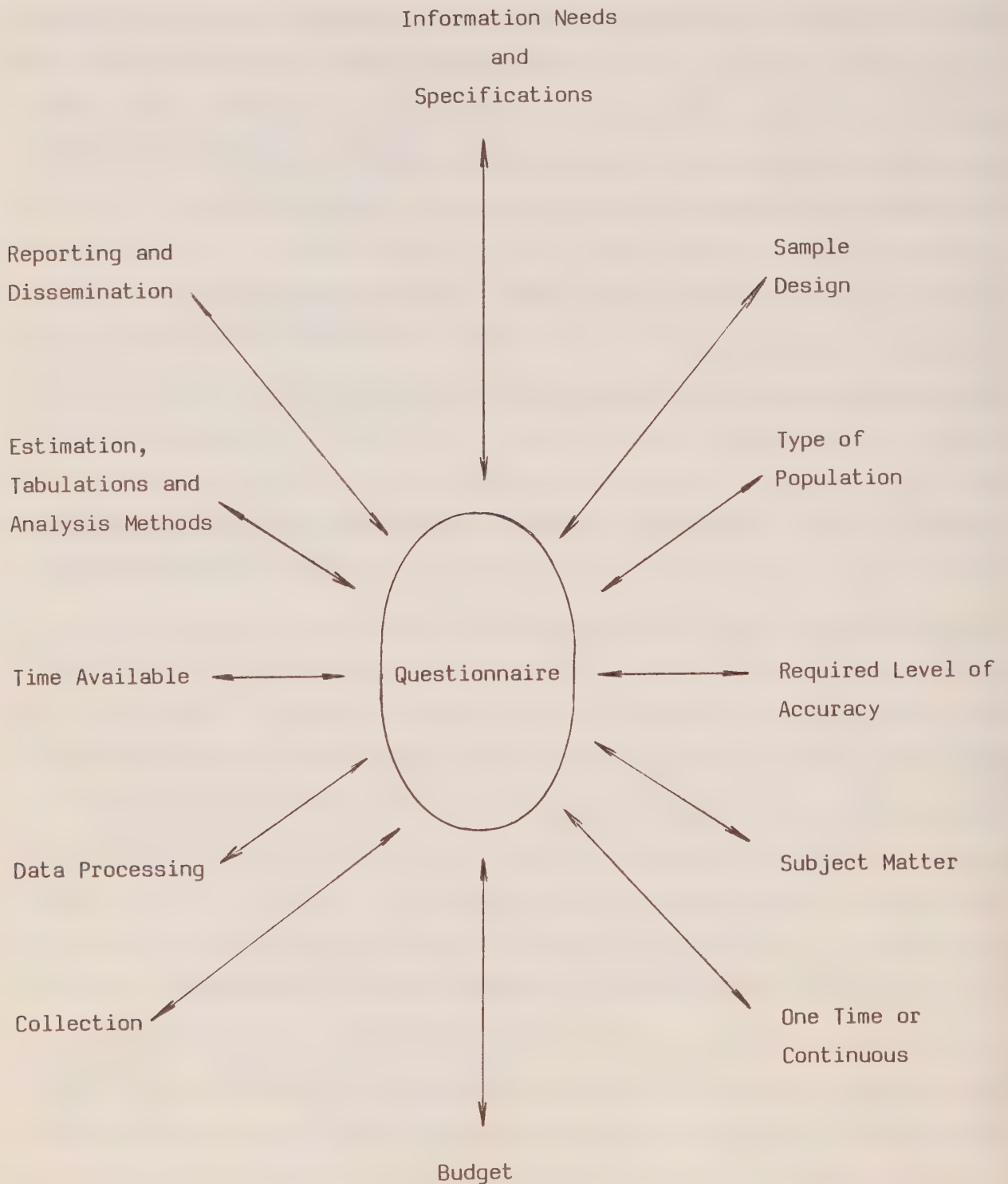
being surveyed, the sample design and sample size, the subject matter of the survey, the interviewing method, the data processing techniques to be used, and the budget and time available.

Figure 1 illustrates the questionnaire's relationships to some of the other elements which make up the total survey design. These interrelationships form a complex network; changes to one component of the design often require changes in several other components as well. Virtually any component of survey design could be placed at the centre of this network, but for the purpose of discussion we have chosen to focus attention on the questionnaire.

Elements such as the type of population, the sample design and the required level of accuracy are closely interrelated with questionnaire design. For example, the heterogeneous nature of many survey populations results in a need for cross-classified data. These needs affect the sample size, the type and degree of stratification, and the reliability of the information. This in turn will affect the questionnaire through the types of questions asked and the level of detail requested. This will further have an effect on the cost and timeliness of the information, the amount of respondent burden, and so on.

The questionnaire design is closely linked to the method of data collection and the survey's subject matter. Each method of data collection, such as personal interviewing, telephone interviewing and mail surveys, creates its own survey conditions which may be more or less appropriate to a given subject matter. These conditions will in turn affect the questionnaire's style of questioning, content, format, length and so on. In personal interviews, for example, it is often possible for the interviewer to collect certain data, such as type of dwelling and sex of respondent, by direct observation rather than questions. In addition, the questionnaire can be designed for the use of flash cards or other visual aids by the interviewer. The element of face-to-face communication is also a powerful motivating factor for the respondent. A personal interview is often the only choice when a complex, long and demanding questionnaire is involved. In telephone interviews, much of the social interaction between interviewer and respondent is lost and the respondent's

Figure 1: Elements Affecting the Questionnaire



co-operation may be affected. The questionnaire must rely entirely on verbal communication for its success, and the subject matter may have to be less demanding. However, with certain sensitive surveys, (e.g. criminal victimization surveys), the extra distance between interviewer and respondent may actually make it easier to answer questions. In mail surveys, the questionnaire itself assumes the role of interviewer. It must introduce the survey, motivate the respondent to co-operate and guide the respondent in completing the interview. It is a particularly demanding role which must be taken into account in designing the questionnaire.

Whether the survey is one-time or continuing also has an effect on questionnaire design. With a continuing survey, there is often more scope for learning from experience and refining the questionnaire over time. Experiments in question wording, programs to monitor response errors, and other methods of evaluating and improving the questionnaire design may only be feasible with a continuing survey. However, the ability to improve a questionnaire must be balanced against the disadvantages of change: for example the inability to make comparisons over time, the necessity to retrain interviewers, and the necessity to change expensive computer software.

In many continuing surveys, such as the Canadian Labour Force Survey, the same respondents are interviewed several times. The questionnaire must take into account the total response burden during the respondent's stay in the survey. The questionnaire may also have to adapt to different collection methods: for example in the LFS the first interview is conducted in person while in urban areas most subsequent interviews are conducted by telephone. Questionnaires designed for continuing surveys must be developed with the longer term view in mind.

The questionnaire is also interrelated with data processing and budgetary concerns. The format of questions, for example open or closed, has direct implications for operations such as coding, data capture, editing and tabulations. The presence of many open-ended questions increases the time and effort during coding operations, and the programs to edit and tabulate the data become more difficult and costly to write and test.

The questionnaire as an operational expression of user needs thus involves the total survey design itself. Survey design is a combination of intricate components, among which the questionnaire plays a central role. The questionnaire neither determines the form of the other components, nor is its form determined by the others. The process of questionnaire design must flow from and be a part of the total survey design process.

#### 4. THE QUESTIONNAIRE AND ERRORS

All survey-taking is subject to errors from various sources, and in recent years non-sampling error has received increasing attention as a major component of the total survey error (see, for example, Anderson et al (1979), Bailar (1976), Hansen, Hurwitz and Bershad (1961), Koch (1973), and Platek and Singh (1980)). The control of non-sampling errors is an integral and vital part of survey design, requiring specific programs for the diagnosis, measurement and prevention of errors. Further, each program will have its own costs and benefits which must be taken into account in the design of controls (Platek and Singh (1980)).

The questionnaire is both an important source of non-sampling errors, and an important part of programs for their prevention and measurement. The scientific development of data collection has lagged behind that of sample design and estimation; improvements in sampling techniques often deal in fractions of a percent while experiments in question wording may reveal variations of 20 percent or more (Payne (1951)). This section discusses the relationship of the questionnaire to a few of the more important sources of non-sampling errors and illustrates the role of the questionnaire in minimizing these errors.

##### 4.1 Non-response errors

Non-response is one important source of non-sampling error. If the characteristics of interest differ from respondents to non-respondents, bias will almost certainly be introduced into the results. Non-response is basically of two types: the "no contact" type, (e.g. no one home, temporarily absent, bad



weather, etc.) and the "refusal" type. The latter may be either a complete non-response or only non-response to some questions. The questionnaire can do little to eliminate the "no contact" type of non-response but it does play an important role in preventing the refusal.

To understand how the questionnaire does this, it is important to first understand why respondents do or do not respond. Many different psychological forces motivate people to respond to surveys, including an interest in the topic, a desire to be helpful, a belief in the importance of the survey, a feeling of duty, or even a belief in their own importance. Other forces influence people to refuse: for example difficulty in understanding questions, fear of strangers, the feeling of one's time being wasted, difficulty in recalling information, and embarrassing or personal questions. All of these forces will have an effect on the questionnaire design through the way in which survey topics are introduced, the question wording, the questionnaire's appearance and length, assurances of confidentiality, and so on. At the same time, these forces interact with the survey's subject matter, the type of population and the data collection method, which in turn influence the design of the questionnaire.

One must also consider the ability of respondents to respond. Unrealistic demands on the respondent's knowledge or memory, the use of overly difficult and technical language, or excessive demands on the respondent's patience are all sources of non-response which have their roots in the questionnaire. It must be said, however, that the patience of respondents often amazes even hardened survey designers. Chinnappa and Wills (1978) describe an interesting study of non-response to the physical measures component of the Canada Health Survey, where respondents were asked to submit to blood pressure tests, skin-fold measurements, exercise tests, and were even asked to donate blood samples.

A more thorough discussion of the causes and treatments of non-response is given in Platek (1980).



#### 4.2 Response errors

Response errors are a second category of non-sampling errors to which the questionnaire is closely related. Response errors can occur anywhere during the question-answer-recording process, and may be either systematic (response bias) or random (response variance).

Questions on sensitive topics, such as amounts and sources of income, use of alcohol and tobacco, illegal activities or mental illness are subject to large response errors. It is often felt, for example, that the respondent may distort the answer to avoid embarrassment or to appear to conform to societal norms (Warwick and Lininger (1975)). Many questionnaire design techniques have been devised to counter this "social desirability bias", including the anonymous questionnaire, the use of projective questioning techniques,<sup>1</sup> or randomized response techniques in which the respondent chooses which of two (or more) questions he answers by the random choice. However, in a recent study which compared questionnaire responses to external criterion information (e.g. official records or test results), Marquis et al (1981) found, rather surprisingly, that for most items which they studied the response bias was almost negligible, but that the response variance was quite large. This conclusion, if supported by other studies, indicates that measuring and reducing response variance may also be important in sensitive topic surveys. This might involve techniques such as reinterviews, internal consistency checks during the interview, and the collection of other information correlated with the variables of interest. This kind of emphasis has direct implications for questionnaire design.

Questions which depend on the respondent to remember events, such as the taking of a trip or the occurrence of a crime, are another source of response errors. Events may be forgotten, or events which occurred before the reference period may be incorrectly included. Bushery (1981), in an experiment

<sup>1</sup> An example of projective questioning might be the sequence:

1. What do you think most people feel about smoking marijuana?
2. How do you yourself feel about it?

The first question asks for the respondent's view of the societal norm and the second asks for his own view.

with the U.S. National Crime Survey, found that victimization rates with a 3-month reference period were much higher than those reported under a 6-month reference period, which were in turn higher than the victimization rates reported with a 12-month reference period. The bias due to recall loss with the longer reference periods was a much more serious source of error than sampling variability. The choice of an appropriate reference period for questions involving recall has been examined in a number of different subject matter areas (Sudman (1980), National Center for Health Statistics (1972)). Bounded recall, where respondents are interviewed at the end of the reference period, or the use of prominent dates (e.g. Christmas) and calendar aids to jog respondents' memories have shown to be of some value in reducing under-reporting (Neter and Waksberg (1965), Ashraf (1975)). With some topics, however, the only possible way to collect the information is to make the questionnaire into a form of diary, where the respondent records the event during, or shortly after, it happens. Questionnaires of this type are used for the Food Expenditure Survey and the Fuel Consumption Survey of Statistics Canada.

Although questions demanding recall and sensitive topics are important sources of response errors, there are many other causes. For example, an important component of response error is that due to the interviewer, the so-called correlated response error. Each interviewer exerts, to some degree, a common influence on all of the respondents in his/her assignment through the way in which the questions are asked, the way in which the respondent's replies are interpreted and recorded, and so on. The contribution of this component of error to the total survey error is directly related to the size of the interviewer's assignment. In telephone surveys, which may have quite large assignments, the correlated component can be a much more serious error than in personal interviews (Groves and Kahn (1979)). In turn, the correlated response error is more serious in personal interviews than in mail surveys or other surveys of the "self enumeration" type. This consideration was a major reason why the Census of Population and Housing has adopted the drop-off-mailback as the standard technique since 1971. The choice of data collection method in turn has a direct influence on the questionnaire design.

Numerous other examples of response errors could be given. They depend on what question is asked, how the interviewer asks it, the way in which the respondent interprets and answers the question, and the way in which the interviewer interprets and records the answer. The interview is a dynamic, interactive process of communication between interviewer and respondent. How it is handled determines whether or not the interview produces the desired information in an accurate and efficient fashion. In the heat of the interview, it is the questionnaire, through its content, question wording, instructions and layout, which must play the major role in controlling the situation.

#### 4.3 Data processing errors

Once the interview is completed, the questionnaire becomes primarily a data processing document. Errors can occur at all phases of processing including coding, data capture, editing, imputation, estimation and tabulation. The way in which the questionnaire was designed will have a significant impact on the number and type of errors at this stage of the survey.

By including data capture codes right on the questionnaire, for example, data capture errors are usually reduced significantly. The data are captured directly from the questionnaire without first being transcribed onto another form. A step beyond this is the Computer Assisted Telephone Interview. The questionnaire is stored in a computer program, which controls the entire interview process. The questions appear one at a time on a video display terminal in front of the interviewer, who then asks the question and types the respondent's reply directly into the computer. The data can be edited immediately and errors corrected while the respondent is still on the telephone. The process also reduces the incidence of questions missed or of incorrect application of skip instructions.

Editing and imputation errors are also closely related to the questionnaire design. Problems of missing or inconsistent data can often be traced back to faulty questionnaire design. The ability to reconstruct or impute for missing values often depends on what concomitant variables were included on the questionnaire and what kind of fail-safe mechanisms were built in. For example,



in a survey which requests information on several detailed components of income, cases where the information is not given or is incorrect can often be salvaged by including a question asking for total income.

Non-response errors, response errors, and data processing errors are a few of the non-sampling errors which are closely linked to the questionnaire and to the other components of the overall survey design. The questionnaire is inevitably a cause of non-sampling error, but it must also go as far as possible in preventing errors. The degree to which the questionnaire succeeds at this task depends largely on the survey designer's knowledge of the various sources of errors and on his skill in integrating the design of the questionnaire with that of the entire survey. Each new survey may present new problems and pitfalls and as such they must be anticipated and taken into account in developing questionnaires.

## 5. CONCLUSION

The preceding sections have illustrated the questionnaire's role as both an expression of the user's information needs and as an important determinant of the quality of survey data. In both roles, the questionnaire is closely linked to all of the components of survey design. The total survey design, and in particular the questionnaire, must try to maximize both the relevance of the data to the user and the accuracy of the data. Successful questionnaire design incorporates both; we must ask the right question, and we must ask it in the right way.

It is important to underline that users' needs and the requirements of accuracy often conflict. The process of questionnaire design involves tradeoffs. A user may have to ask a simpler question than he would like simply for the respondent to be capable of answering. On the other hand, the questionnaire designer should not avoid asking complex questions simply because the answers may contain errors.

Questionnaire development is not simply a laboratory process. Although guidelines exist and research is possible, the skill of questionnaire design is learned to a large extent by practical experience and by trial and error. It is learned through discussions with users, interviewers and respondents. Questionnaire design is undoubtedly an interactive process which cannot be carried out in isolation and independent of other factors in survey development. It interrelates with them and, in fact, it forms an integral part of the total survey design.

#### REFERENCES

- [1] Andersen, R., Kasper, J., Frankel, M.R. and Associates (1979), Total Survey Error, Jossey-Bass Publishers, San Francisco.
- [2] Ashraf, A. (1975), "The Methodology of the Canadian Travel Survey", Survey Methodology, Vol. 1, No. 2, 208-227.
- [3] Bailar, B. (1976), "Some Sources of Error and Their Effect on Census Statistics", Demography, Vol. 13, No. 2, 273-286.
- [4] Bushery, J.M. (1981), "Recall Biases for Different Reference Periods in the National Crime Survey", paper presented at the 141st Annual Meeting of the American Statistical Association, Detroit.
- [5] Carson, E.M. (1973), Questionnaire Design: Some Principles and Related Topics, Statistics Canada internal report.
- [6] Chinnappa, B.N. and Wills, B. (1978), "A Study of Refusal Rates to the Physical Measures Component of the Canada Health Survey", Survey Methodology, Vol. 4, No. 2, 100-114.
- [7] Dillman, D.A. (1978), Mail and Telephone Surveys: The Total Design Method, Wiley, New York.



- [8] Fellegi, I. (1979), "Data, Statistics, Information - Some Issues of the Canadian Social Statistics Scene", *Survey Methodology*, Vol. 5, No. 2, 130-161.
- [9] Groves, R.M. and Kahn, R.L. (1979), Surveys by Telephone, Academic Press, New York.
- [10] Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", *Bulletin of the International Statistical Institute*, Vol. 38, Part II, 359-374.
- [11] Jabine, T.B. (1981), Guidelines and Recommendations for Experimental and Pilot Survey Activities in Connection with the Inter-American Household Survey Program, Inter-American Statistical Institute report 7679 a - 5/7/81 - 100, Washington, D.C.
- [12] Kahn, R.L. and Cannell, C.F. (1957), The Dynamics of Interviewing, Wiley, New York.
- [13] Koch, G. (1973), "An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means", *Journal of the American Statistical Association*, Vol. 68, No. 344, 906-913.
- [14] Marquis, K.H., Marquis, M.S. and Polich, J.M. (1981), "Survey Response Errors for Sensitive Topics: The Problem is Noise Rather than Bias", paper presented at the 141st Annual Meeting of the American Statistical Association, Detroit.
- [15] National Center for Health Statistics (1972), "Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents", *Vital and Health Statistics*, Series 2, No. 50, U.S. Government Printing Office, Washington, D.C.

- [16] Neter, J. and Waksberg, J. (1965), Response Errors in Collection of Expenditures Data by Household Interviews: an Experimental Study, Bureau of the Census Technical Paper No. 11, U.S. Government Printing Office, Washington, D.C.
- [17] Payne, S.L. (1951), The Art of Asking Questions, Princeton University Press, Princeton, N.J.
- [18] Platek, R. (1980), "Causes of Incomplete Data, Adjustments and Effects", Survey Methodology, Vol. 6, No. 2, 93-132.
- [19] Platek, R. and Singh, M.P. (1981), "Cost Benefit Analysis of Controls in Surveys", Current Topics in Survey Sampling, Academic Press, New York.
- [20] Shannon, C.E. and Weaver, W. (1949), The Mathematical Theory of Communication, University of Illinois Press, Urbana.
- [21] Statistics Canada (1979), Basic Questionnaire Design, Second edition, Ottawa.
- [22] Sudman, S. (1980), "Reducing Response Errors in Surveys", Statistician, Vol. 29, 237-273.
- [23] Warwick, D.P., and Lininger, C.A. (1975), The Sample Survey: Theory and Practice, McGraw - Hill, New York.

## EVALUATION OF SMALL AREA ESTIMATION TECHNIQUES FOR THE CANADIAN LABOUR FORCE SURVEY<sup>1</sup>

J.D. Drew, M.P. Singh, G.H. Choudhry<sup>2</sup>

Estimates from sample surveys are sometimes required for domains whose boundaries do not coincide with those of design strata. Taking the Canadian Labour Force Survey as an example of a survey utilizing a clustered sample design, some alternative small area estimation techniques available in the literature are evaluated empirically including synthetic, domain (simple and post-stratified) and composite estimators which are linear combinations of synthetic and post-stratified domain estimators. A sample dependent estimator which attaches weight to the post-stratified domain estimate depending on the amount of sample in the domain is proposed and its performance is also evaluated.

### 1. INTRODUCTION

With increasing emphasis on planning, administering and monitoring social and fiscal programs at local levels, there has been demand for more and good quality data at these levels from various municipal, provincial and federal government departments as well as from private institutions. The type of data required ranges from simple population counts to complex socio-economic variables such as employment, unemployment, income, housing, poverty indices, health conditions and facilities etc. However, until recently not much attention had been paid to the development of sound statistical estimation techniques for small area data, with the notable exception of statistical demographers who for some time have been investigating the particular problem of small area population estimates, and who have identified several competing methods based on the use of administrative data and other sources.

<sup>1</sup> Presented at the Annual meetings of the American Statistical Association, Cincinnati, August 1982

<sup>2</sup> J.D. Drew, M.P. Singh and G.H. Choudhry, Census and Household Survey Methods Division, Statistics Canada.

A comprehensive review of existing small area (domain) estimation techniques along with their limitations is given by Purcell and Kish (1979). From the research done to date it is clear that there is not a unique best solution to the small area estimation problem. The choice of a particular method for small area estimation will depend on the data needs and on the richness and availability of data sources, which differ from country to country, and within countries from one subject matter to another. Therefore, the classification of the type of small areas (domains) and examination of the data sources available in a particular context, followed by thorough investigation of the alternative small area estimation techniques for given situations, seems to be the most appropriate approach to development of small area data. In this context, we shall use the following classification of domains suggested by Purcell and Kish (1979) and point out the type of domain to which developments in this articles primarily refer.

- (a) Planned domains - for which separate samples have been planned, designed, and selected. In the Canadian context, such domains for example may be economic or planning regions within a province or the province itself.
- (b) Cross Classes - which cut across the sample design and the sample units (may also be referred to as characteristic domains); e.g., age/sex, occupation, industry.
- (c) Unplanned Domains - that have not been distinguished at the time of sample design and thus may cut across the design strata or the primary sampling units (PSU's) within the strata. Examples of these in the Canadian context include Federal Electoral Districts, and Census Divisions or subdivisions, counties and manpower planning regions.

It should be noted that both types (a) and (c) refer to areal domains.

We consider this distinction of the domains into the above types important since the form of the estimator as well as its efficiency would depend upon the particular type of application. As pointed out by Purcell and Kish most of the developments in small area estimation techniques in the United States



and elsewhere have concentrated on the domains of types (a) and (b). In Canada however, type (a) and (b) domains are not so problematic due to the type of design and the sizes of the national surveys, and the main emphasis has been on the data for the domains of type (c), with the possible exception of population counts using symptomatic data.

Investigations into the application and evaluation of small area estimation techniques for variables other than population started with the publication of synthetic estimates from the National Center for Health Statistics (1968). Since then a series of investigations (Gonzalez (1973), Gonzalez and Waksberg (1973), Schaible, Brock and Schnack (1977), Gonzalez and Hoza (1978) and others) have been carried out using data from the Current Population Survey in the application and evaluation of a particular synthetic estimator. Using a synthetic estimator whose form is different, studies were carried out by Purcell and Linacre (1976) aimed at production of estimates for Census divisions in Australia and by Ghangurde and Singh (1976, 1977, 1978) in the evaluation of synthetic estimates in the context of Canadian Labour Force Survey (LFS).

As remarked by Purcell and Kish (1979), the nature of the design in relation to the domains of interest has an important role to play in the choice of an estimator. The estimators considered in this article are geared to the Canadian LFS where the domains are unplanned domains (type c) and are of a size such that, had they been planned domains (type a), the reliability of regular unbiased survey estimates would be satisfactory without having to resort to small areas estimation techniques. Also in the LFS, primary sampling units are small (populations from 2,000 - 5,000) relative to the sizes of the domains of interest. This differs from the situation in the United States where the sizes of primary sampling units for most of the large scale surveys are larger, comparable in size to the small areas for which the estimates are desired.

In this article estimators are evaluated in the context of producing Census Division level estimates from the Labour Force Survey, using data from the 1971 and 1976 Censuses of Population and Housing in an auxiliary fashion. In



addition to synthetic estimators, we evaluate post-stratified domain estimators which were considered earlier by Singh and Tessier (1976), and composite estimators which are linear combinations of the synthetic and the post-stratified domain estimators, similar to those considered by Schaible (1979) and Schaible, Brock and Schnack (1977). Also we propose and evaluate a new estimator which we call a sample dependent estimator, which is of the same form as the composite estimator, except the weight given to the synthetic component is a decreasing function of the amount of sample falling into the domain upto a critical point after which the estimator relies totally on the post-stratified domain component. Efficiencies of the small area estimators relative to the direct (or simple domain) estimator for the characteristics employed and unemployed were obtained in an empirical (Monte Carlo) study in which the LFS design was simulated using census data. The situations where both the design and the auxiliary information are up-to-date and where both are out-of-date were considered. We have also evaluated the bias of synthetic estimators for the characteristics employed and unemployed for Federal Electoral Districts.

## 2. DESCRIPTION OF ESTIMATION PROCEDURE

Consider a finite population consisting of  $N$  units, (e.g. households or dwellings in household surveys), divided into  $L$  design strata labelled  $1, 2, \dots, h, \dots, L$ . The stratification has been carried out on the basis of geographic and/or certain socio-economic characteristics, and the sample allocation ensures certain precision for estimates from individual strata. The problem considered is that of estimating the total of an  $x$ -variate for all those units belonging to an unplanned areal domain (type c). We denote by 'a' the set of units belonging to the small area or domain of interest, thus the parameter to be estimated is the total of the  $x$ -variable in the domain 'a', which we denote by  $X_a$ .

Let  $a_h$  be the set of those units belonging to the domain which are in stratum  $h$ , then

$$a = \bigcup_{h=1}^L a_h. \quad (2.1)$$

In practice the domain 'a' will have a non-null intersection with a certain number of design strata and if we denote by  $h_{\sim}$  the set of such strata, then we have

$$a = \bigcup_{h \in h_{\sim}} a_h. \quad (2.2)$$

The particular design under consideration follows a multi-stage clustered sample design which is self-weighting within each stratum with weight  $w_h$  for stratum h.

For a particular given sample we can obtain the quantities:

$$t_h = \text{sample total of } x\text{-variate in stratum } h,$$

and

$${}_a t_h = \text{sample total of } x\text{-variate in } a_h$$

for  $h=1, 2, \dots, L$ . Note that  ${}_a t_h = 0$  for  $h \notin h_{\sim}$ . Then the direct (or also referred to as design based or simple domain) estimator for the total of x-variate for those units in 'a' say  $\hat{\chi}_a$ , is given by:

$$\hat{\chi}_a = \sum_{h \in h_{\sim}} w_h \cdot {}_a t_h. \quad (2.3)$$

It should be noted that the direct estimator (2.3) does not utilize any auxiliary information - all it requires is the identification of those sampled units which belong to the domain. Due to the clustered nature of the design,

the sample falling in the domain may on occasion be very small or non-existent, generally resulting in high variance for this estimator.

The other estimators in this section rely in different fashions on auxiliary information for a variable  $y$ , which is often taken as the count of persons by population sub-groups (defined on the basis of age/sex etc.) from a recent census. These estimators are:

- 1) Post-stratified domain
- 2) Synthetic
- 3) Composite
- 4) Sample Dependent

Additionally estimators (2) - (4) rely to differing degrees on sample external to the domain.

For each of the above estimators, the adjustments based on the auxiliary information can be made either by applying separate adjustments to each stratum intersecting the domain, or by applying an overall adjustment for all strata intersecting the domain. Thus the estimators will be further classified as separate or combined depending on the level at which the adjustment is made. These estimators are denoted by  $\hat{\chi}_{uv}$ , where  $u$  is the level of adjustment with values:

$u = s$  : separate  
 $= c$  : combined

and  $v$  is the type of estimator taking the following values:

$v = p$  : post-stratified domain  
 $= s$  : synthetic  
 $= c$  : composite  
 $= d$  : sample dependent

For example,  $\hat{\chi}_{cs}$  denotes the combined synthetic estimator, etc.

## 2.1 Post-Stratified Domain Estimator

Define

$Y_{hg}$  = total of the auxiliary y-variable for population sub-group g in group g in stratum h, and

${}_aY_{hg}$  = total of the auxiliary y-variable for population sub-group g in  $a_h$ .

Further let  $\tilde{{}_aY_{hg}}$  be an unbiased estimated of  ${}_aY_{hg}$  which would be formed analogously to the direct estimate defined in (2.1), except the characteristic being estimated in this case would be the auxiliary y-variable whose value is known for the set of sampled units (s) at some stage of sampling (whereas (2.1) is defined on the x-variate for the sample of ultimate units). In practice provided auxiliary y-variable information is available for them, sampling units at any stage down to the penultimate stage could be used.

Then the separate post-stratified domain estimator (for which adjustments are applied at the stratum level) is:

$$\hat{X}_{a\ sp} = \sum_g \sum_{h \in \tilde{a}_h} (W_h \cdot t_{a\ hg}) \frac{{}_aY_{hg}}{\tilde{{}_aY_{hg}}} \quad (2.4)$$

where  $t_{a\ hg}$  is the sample total of the x-variate for population sub-group g in the intersection of domain 'a' and stratum h.

Similarly the combined post-stratified domain estimator (for which adjustments are applied at the domain level) is:

$$\hat{X}_{cp} = \sum_g \sum_{h \in \tilde{h}} (W_h \cdot t_{hg}) \frac{\sum_{h \in \tilde{h}} a_{hg} Y_{hg}}{\sum_{h \in \tilde{h}} Y_{hg}} \quad (2.5)$$

The post-stratified domain estimator is unbiased except for the effect of ratio estimation bias, provided  $Y_{hg}$  is obtained at the same time as  $a_{hg}$  and using the same source such as census.

Estimators of the above type have been considered earlier by Singh and Tessier (1976) with a different choice of post-strata.

## 2.2 Synthetic Estimators

We consider separate and combined synthetic estimators defined respectively as follows:

$$\hat{X}_{ss} = \sum_g \sum_{h \in \tilde{h}} (W_h \cdot t_{hg}) \frac{a_{hg} Y_{hg}}{Y_{hg}} \quad (2.6)$$

$$\hat{X}_{cs} = \sum_g \sum_{h \in \tilde{h}} (W_h \cdot t_{hg}) \frac{\sum_{h \in \tilde{h}} a_{hg} Y_{hg}}{\sum_{h \in \tilde{h}} Y_{hg}}, \quad (2.7)$$

where  $t_{hg}$  is the sample total for the x-variable for population sub-group g in stratum h.

The above synthetic estimator has been considered by Purcell and Linacre (1976) and also by Ghangurde and Singh, (1976, 1977, 1978) who developed expressions for its variance and bias and evaluated the estimator using census data and a super-population model. A different form of synthetic estimator was proposed earlier by the National Centre for Health Statistics (1968) and investigated by Gonzalez (1973), Gonzalez and Waksberg (1973) and Gonzalez and Hoza (1975, 1978) using data from the Current Population Survey.



The difference between the synthetic and post-stratified domain estimators can be readily seen by comparing (2.4) and (2.6). The post-stratified domain estimator uses only the sample falling into the domain (i.e.,  $t_{hg}$ ) and the adjustment factor is the ratio of the true to the estimated values for the y-variable for the domain and hence can take on values greater than or less than 1 (its expected value being unity). On the other hand the synthetic estimator uses the estimate from entire strata intersected by the domain (i.e.  $w_h \cdot t_{hg}$  for  $h \in \tilde{h}$ ) which is then deflated by adjustment factors specific to population subgroups. (i.e. the ratio of the y-variable for the domain to the y-variable for the entire stratum).

The synthetic estimator will suffer from bias depending on the degree of departure from the assumption of homogeneity for the x-variate between the domain and the larger area, namely  $h_{\tilde{h}}$ , withing sub-groups of the y-variable. In defining the above synthetic estimator, the larger area was restricted to those strata which form part of the domain as it was believed that such a choice would lead to less bias. In general however,  $h_{\tilde{h}}$  need not be so restricted but it may include other neighbouring areas which are believed to satisfy the homogeneity assumption. Bias and mean square error of such estimators have been reported by some of the earlier referenced authors.

### 2.3 Composite Estimators

A composite estimator using the direct estimator and the synthetic estimator as the two components was suggested by Royall (1973) and others, and has been studied by Schaible (1978). Such an estimator minimises the chances of extreme situations (both in terms of bias and mean square error) and therefore may be preferred over either of its components. Synthetic estimators have a low variance by virtue of their use of data from a larger area to derive estimates for small area (domain), but for the same reason this introduces bias which could be quite large if as noted earlier, the assumption of homogeneity is not satisfied. On the other hand the simple domain estimator, which is unbiased, may have large variance particularly if the sample falling in the domain is very small. Empirical evidence of such relative performances of synthetic and direct estimators are available from Gozalez and Waksberg (1975)

Schaible, Brock and Schnack (1977), and Ghangurde and Singh (1977). The composite estimator considered here is obtained by replacing the direct estimator (2.3) by the post-stratified domain estimator which may be slightly biased but is generally more efficient than the direct estimator.

The two types of composite estimators: namely, separate and combined are formed as linear combinations of the corresponding post-stratified domain and synthetic estimators; viz,

$$\hat{X}_{sc} = \alpha_1 \hat{X}_{sp} + (1 - \alpha_1) \hat{X}_{ss} \quad (2.8)$$

and

$$\hat{X}_{cc} = \alpha_2 \hat{X}_{cp} + (1 - \alpha_2) \hat{X}_{cs} \quad (2.9)$$

The optimum values for  $\alpha_1$  and  $\alpha_2$  for minimum mse's are given by

$$\alpha_1^* = \frac{\text{mse} [\hat{X}_{ss}] - E [\hat{X}_{ss} - a^X] [\hat{X}_{sp} - a^X]}{\text{mse} [\hat{X}_{ss}] + \text{mse} [\hat{X}_{sp}] - 2 E [\hat{X}_{ss} - a^X] [\hat{X}_{sp} - a^X]} \quad (2.10)$$

and a similar expression for  $\alpha_2^*$ .

Further, neglecting the covariance term in (2.10) under the assumption that this term will be small relative to  $\text{mse} [\hat{X}_{ss}]$  and  $\text{mse} [\hat{X}_{sp}]$ , then the optimal weight  $\alpha_1^*$  can be approximated by

$$\alpha_1^{**} = \frac{\text{mse} [\hat{X}_{ss}]}{\text{mse} [\hat{X}_{ss}] + \text{mse} [\hat{X}_{sp}]} \quad (2.11)$$

with a similar expression for  $\alpha_2^{**}$ , which was the approach to defining weights followed by Schaible (1978).

## 2.4 Sample Dependent Estimators

In practice the true values of  $\alpha_1^*$  (or  $\alpha_2^*$ ) used as the weight in the composite estimator will not be available as they involve population variances and covariances, which would have to be estimated from the sample. Further calculation of the covariance term in (2.10), in particular, may be quite complex and thus one may have to resort to an approximate value  $\alpha_1^{**}$  (or  $\alpha_2^{**}$ ) which would require simply the estimated mse's of the two component estimators or an estimate of the ratio of the two mse's. In either case there estimates would introduce a certain amount of instability in the weight used, thus affecting the performance of the composite estimator.

The sample dependent estimator (Drew and Choudhry, (1979)) which is a particular case of a composite estimator, depends on the outcome of the given sample and is quite simple to compute. It is constructed using the result that the performance of the post-stratified domain estimator depends upon the proportion of the sample falling in the domain. If the proportion of the sample within the domain is 'reasonably large' then the sample dependent estimator is the same as the post-stratified domain estimator, otherwise it becomes a composite estimator with gradual increasing reliance (in the sense of increasing weight) on the synthetic estimator as the size of the sample in the domain decreases. Thus the separate sample dependent estimator (i.e., constructed at the stratum level) is given by

$$\hat{X}_{a\ sd} = \sum_g \sum_{h \in \tilde{a}} \left[ \delta_{hg} \frac{W_h}{Y_{a\ hg}} \cdot t_{a\ hg} \frac{Y_{hg}}{\tilde{Y}_{a\ hg}} + (1 - \delta_{hg}) \frac{W_h}{Y_{a\ hg}} \cdot t_{a\ hg} \frac{Y_{hg}}{\tilde{Y}_{a\ hg}} \right] \quad (2.12)$$

where

$$\delta_{hg} = 1, \text{ if } \tilde{Y}_{a\ hg} / Y_{a\ hg} \geq K_0,$$

$$= \frac{1}{K_0} \frac{\tilde{Y}_{a\ hg}}{Y_{a\ hg}}, \quad \text{otherwise.}$$

Similarly the combined sample dependent estimator (i.e. constructed at the domain level) is given by

$$\begin{aligned} \hat{X}_{cd} = & \sum_g \left[ \delta \left( \sum_{h \in \tilde{a}} W_{hg} \cdot t_{hg} \right) \frac{\sum_{h \in \tilde{a}} Y_{hg}}{\sum_{h \in \tilde{a}} W_{hg}} \right. \\ & \left. + (1 - \delta) \left( \sum_{h \in \tilde{a}} W_{hg} \cdot t_{hg} \right) \frac{\sum_{h \in \tilde{a}} Y_{hg}}{\sum_{h \in \tilde{a}} W_{hg}} \right] \end{aligned} \quad (2.13)$$

where

$$\begin{aligned} \delta_g &= 1, \text{ if } \frac{\sum_{h \in \tilde{a}} Y_{hg}}{\sum_{h \in \tilde{a}} W_{hg}} \geq K_o \\ &= \frac{1}{K_o} \frac{\sum_{h \in \tilde{a}} Y_{hg}}{\sum_{h \in \tilde{a}} W_{hg}}, \quad \text{otherwise} \end{aligned}$$

The ratios

$$\frac{\sum_{h \in \tilde{a}} Y_{hg}}{\sum_{h \in \tilde{a}} W_{hg}} \quad \text{and} \quad \frac{\sum_{h \in \tilde{a}} Y_{hg}}{\sum_{h \in \tilde{a}} W_{hg}}$$

indicate the over- or under-representation of the population sub-group at the individual stratum or domain level with respect to auxiliary information for the y-variable, conditional upon the selected sample.

Values of ratios greater than or equal to 1 signify that, conditional on the given sample (s), the representation of the population sub-groups for the auxiliary y-variable is better than or as good as its unconditional representation had the domain been sampled independently at the same rate as the stratum.

The value of  $K_o$  may be appropriately chosen. In this study the efficiency of sample dependent estimator has been investigated for two specific values of  $K_o$  namely 1.0 and 0.5.

Holt, Smith and Tomberlin (1979) under the prediction approach derived an estimator (which relies on synthetic and direct estimates) where the weight attached to the direct component depends only on the sample falling into the domain. Sarndal (1981) proposed an alternative estimator in which the weight attached to the direct component depends on the sample in the domain relative to the sample in the larger area.

### 3. DESCRIPTION OF THE EMPIRICAL STUDY

#### 3.1 Simulation of the LFS Design

The LFS follows a multi-stage area sampling design (see Platek and Singh, (1976)). Within each of the 10 provinces of Canada, two principal area types are identified - the Self-Representing Units (SRU's) which correspond to cities generally of 15,000 or more population, and the Non Self-Representing Units (NSRU's) which correspond to smaller urban centers and rural areas. In the SRU's, cities are divided into compact areal strata with populations of 15,000 each, within which a two stage sample of clusters (similar to blocks) and dwelling is selected.

In NSRU's, Economic Regions, of which there are from 1-10 per province, form the starting point. These are stratified into 1-5 strata with populations from 30,000 to 80,000 using census data for 7 broad industry classifications. Within strata, primary sampling units (PSU's) from 2,000 - 5,000 in population are formed. The second stage in the rural portions of PSU's corresponds to 1971 Census Enumeration Areas (i.e., EA's), with populations of roughly 500, whereas in urban portions all urban centers are selected with certainty. The last two stages correspond to clusters and dwellings.

In simulating the LFS design two cases were examined: (i) the case where both the sample design and the auxiliary information are up-to-date, and (ii) the case where both are out-of-date.



For (i), the sample design, the auxiliary information, and the study variables were all based on 1971 census data. Counts of persons (15+) cross-classified by age/sex, and Labour Force status were retrieved at the EA level. In NSRU's, for each replication in the Monte Carlo study independent samples of primaries and secondaries were selected based on census population or dwelling counts. Within rural EA's and urban centers, the final two stages of sampling were simulated by random samples of persons. In SRU's, EA's comprising the areal strata were known, but there after the LFS design was independent of the census. Hence for the purposes of the study, EA's were randomly partitioned into 'clusters' having a size distribution corresponding to that for LFS clusters. For each replication, a sample of 'cluster' and a random sample of persons within were selected.

### 3.2 Choice of Population Sub-Groups

The estimators defined in section 2 utilize auxiliary information for population sub-groups. Since the LFS is redesigned only decennially, it would be desirable to base the population sub-groups on information collected in the mid-decade as well as decennial census, so that the auxiliary information could be updated mid-way through the life of the survey. This ruled out such variables as industry or occupation, leaving various cross-classifications of basic demographic variables as the possible choices for population sub-groups.

For the variables marital status, age and sex, the Automatic Interaction Detection (AID) procedure, due to Sonquist and Margan (1964) was used on a sample of census data from across Canada to derive optimal population sub-groups, separately for each Labour Force characteristic. Results of the AID analysis showed that for unemployed, no population sub-groups accounted for more than 2% of the variation, while for the characteristics employed and not in Labour Force the following sub-groups accounted for approximately 25% of the variation: (i) age 15-16 and 65+; (ii) age 17-64, sex female; (iii) age 17-64, sex male. Further splitting of these sub-groups did not result in significant additional gains.

In addition to estimators based on the above population sub-groups, estimators based on total population 15+, and on dwelling counts were also considered. Dwelling count data were included due to the possibilities which exist for up-to-date dwelling information being available intercensally at the required level of detail. It might be noted that the estimators using population 15+ and dwelling counts are both special cases of the general formulation where the number of population sub-groups equals 1.

### 3.3 Evaluation of Efficiency of Small Area Estimators

In the Monte Carlo study, we have considered 16 Census Divisions (CD's) and 11 Federal Electoral Districts (FED's) in the province of Nova Scotia and 7 FED's from elsewhere in Canada. (There are altogether 18 CD's in the province of Nova Scotia, but two of 18 CD's correspond to complete LFS strata and therefore were omitted from the study). Due to the multi-stage nature of the design and larger number of domains in the study, the computational costs involved were high and it was decided to use only 100 replications.

Census Divisions and Federal Electoral Districts, it should be noted, comprise networks of geo-statistical and geo-political areas respectively across Canada. There are approximately 300 of each, with the populations of Federal Electoral Districts being fairly uniform in the range 80,000 to 120,000, while those of Census Divisions, which often correspond to local levels of government or counties, vary greatly.

We have reported results only for the 16 Census Divisions in Nova Scotia. Results were similar for other unplanned domains considered.

If we let  $\hat{a}X_{m(r)}$  be the estimate of total  $aX$  (i.e. the total for the x-variable for the domain 'a') for the r'th replicate, for small area estimation method m, then the average mean square error for the method m over the 16 domains in the study was calculated as:

$$\text{Avg mse (m)} = \frac{1}{16} \sum_a \sum_{r=1}^{100} (\hat{a}X_{m(r)} - aX)^2 / 100 \quad (3.1)$$

The efficiency of the small area estimator ( $m$ ) relative to the direct estimator, say method  $m_0$  was obtained as:

$$\text{Eff } (m \text{ vs } m_0) = \frac{\text{Avg mse } (m_0)}{\text{Avg mse } (m)}. \quad (3.2)$$

### 3.4 Evaluation of Bias of Synthetic Estimators

Since the composition of the LFS frame and the Federal Electoral Districts were known for all of Canada in terms of both 1971 and 1976 census units, it was possible to compute exact biases of the synthetic estimators based on census data. The following cases were considered: (i) design and auxiliary information up-to-date (in which case the design, adjustment factors and x-variables were all based on the 1971 census); and (ii) design and auxiliary information out-of-date (in which case the design and adjustment factors were based on the 1971 census, but the x-variables were based on the 1976 census).

Let  $aB_{ss}$  and  $aB_{cs}$  denote the biases of the separate and combined synthetic estimates for unplanned domain 'a', then we have

$$B_{a \text{ ss}} = \sum_g \sum_{h \in \tilde{h}} (X_{hg} \frac{Y_{a \text{ hg}}}{Y_{hg}} - X_{hg}) \quad (3.3)$$

and

$$B_{a \text{ cs}} = \sum_g ( \sum_{h \in \tilde{h}} X_{hg} \frac{\sum_{h \in \tilde{h}} Y_{a \text{ hg}}}{\sum_{h \in \tilde{h}} Y_{hg}} - \sum_{h \in \tilde{h}} X_{hg} ) \quad (3.4)$$

where  $aY_{hg}$  and  $Y_{hg}$  are defined as in section 2, and where  $X_{hg}$  and  $aX_{hg}$  are similarly defined for the x-variable (based on the census).

Relative absolute biases at the province level were obtained by summing the absolute biases over individual FED's and dividing by the provincial total for the x-variable.

#### 4. ANALYSIS OF RESULTS

##### 4.1 Efficiency considerations: Auxiliary Information up-to-date

In this part of the empirical (Monte Carlo) study, data used for simulation of the design and the auxiliary variables used in estimation refer to the same period as those of the study variable; i.e., to the 1971 census. Efficiencies of the four small area estimators are presented relative to the direct estimator in Table 1, for separate and combined levels of construction, and for each of the following auxiliary variables - dwellings, total population (15+), and population by age/sex groups. Census Divisions in the province of Nova Scotia whose populations range from 3,885 to 39,260 were used as the unplanned domains (type c) for the purpose of the study. The following observations can be made:

- (i) Separate vs Combined Estimator: The level of construction of estimator does not have much impact on the efficiencies of synthetic estimators for both the characteristics employed and unemployed. For the post-stratified domain estimator for employed, however, the combined form is approximately twice as efficient as the separate. This is likely due to the effect of the clustering in the sample design being more accentuated with the separate estimator.

Since the post-stratified domain estimator was less efficient in its separate form, a similar result was anticipated for the composite estimator and hence, only the combined composite estimator was considered. On the other hand, the separate form of the sample dependent estimator was found to rely slightly more on the synthetic component, leaving the efficiencies unaffected by the level of construction.



- (ii) Effect of Auxiliary Information: The performance of population by age/sex as an auxiliary variable is uniformly superior, although only marginally so, to the total (15+) population for all four estimators using auxiliary information. Further, both these variables out-perform the dwelling count as an auxiliary variable.

In actual survey situations, the choice of population by age/sex as the auxiliary variable may be desirable also from the point of view of correcting estimates for biases due to non-response and undercoverage as both factors may be dependent on age and sex.

- (iii) Comparison among the estimators: For unemployed, performance of composite estimator with optimum  $\alpha_2^*$  chosen for the characteristic unemployed is marginally superior to the other estimators irrespective of the level of construction, and the choice of auxiliary variable does not seem to have appreciable impact on any of the estimators. For employed, the situation is not that clear, however the sample dependent estimator shows an edge over other estimators and particularly so with population by age/sex as the auxiliary variable.

#### 4.2 Efficiency Considerations: Auxiliary information out-of-date

In this part of the study whereas the design and auxiliary information were based on 1971 census results, the study variable was based on the 1976 census. As can be seen from table 2, although for unemployed the use of small area estimation techniques showed larger gains relative to the direct estimator (than in the up-to-date case), considerably smaller gains were observed for employed, which would likely be due to the reduced correlation between the study variable and the auxiliary information as both design and auxiliary information become out-of-date. Also in this case, the efficiency of the synthetic estimator is higher for both of the characteristics measured.

#### 4.3 Consideration of Bias

Given that the post-stratified domain estimator will generally have negligible bias, the bias of both the composite and sample dependent estimators would



generally be smaller than that of the synthetic estimator, i.e. stemming only from the degree of reliance on the synthetic component. Hence the bias of synthetic estimator was investigated in detail. Using the total population (15+) as the auxiliary variable, the relative bias for the characteristics employed and unemployed were computed and are given in Table 3 for the ten provinces. These biases refer to the case where the unplanned domains are Federal Electoral Districts and the study variables are based on 1976 census data, while the survey design and adjustment factors (synthetic weights) are based on the 1971 census. Biases were also computed using age/sex sub-groups as the auxiliary variable and were found to follow similar trends while being marginally smaller. It is observed from this table; with the exception of the two smaller provinces, namely P.E.I. (for unemployed) and N.B. (for employed), that the relative bias of separate synthetic estimator is smaller than that of the combined synthetic estimator for both the characteristics under study. This confirms the intuitive feeling that the higher the level at which synthetic estimator is constructed, the higher would be the resultant bias in general, due to weakening of the assumption of homogeneity.

Biases were also computed for the case when both the study variable and the auxiliary information referred to the 1971 census. Biases for this case while slightly lower, followed similar trends to those in Table 3.

While the bias of the synthetic estimator was fairly small on average, it can be observed from Table 4 that it exceeded 10% in 13 and 19 (out of 279) FED's when the auxiliary information was up-to-date and out-of-date respectively. Further, in about half the instances for which the bias exceeded 10% for the up-to-date case the bias also exceeded 10% for the later time period when the auxiliary information was out-of-date. This suggests that for domains with a known high bias at the time to which the auxiliary information refers, less use should be made of the synthetic estimator. For instance, with the sample dependent estimator the value of  $K_0$  could be set lower in such cases. However there is still the danger of bias in the synthetic estimator from category (ii) type cases in Table 4 which cannot be identified when deriving current estimates during the intercensal period.

#### 4.4 Efficiency vs Bias in Overall Choice of Estimator

The synthetic estimator is generally highly biased and at the same time highly efficient. Therefore, in the search for a reasonable estimator for small areas, the question is to what extent one can reduce the effect of the synthetic estimator's bias, without sacrificing too much on its efficiency, in order to obtain a 'reasonable level of confidence' in the final estimate. At the same time it is also important to determine the reliance on the synthetic estimator without introducing too many computational complexities. Looking from this perspective in the context of the Labour Force Survey, one should strive for small area estimators whose performance for unplanned domains is comparable to that of simple survey estimates for planned domains, and amongst estimators meeting this criterion, more emphasis should be on reducing bias than on improving efficiency, especially if the differences in efficiencies are minor.

Average variances of the unbiased design estimator for the planned domains (say  $\hat{X}$ ), comparable in size to the unplanned domains were obtained analogously to the average mse defined in (3.1). The efficiencies of the synthetic, composite and sample dependent estimators relative to the usual survey estimate for the planned domain i.e.  $X$  were also obtained. These efficiencies ranged from 1.08 to 1.17 for unemployed, and 1.22 to 1.47 for employed, hence all three estimators meet the above mentioned criterion. Since the sample dependent estimator makes use of the synthetic estimators whenever there is not 'sufficient' sample in the domain, its bias would depend upon the weight attached to the synthetic estimator component and this can be controlled by a proper choice of  $K_0$ . Table 5 presents the  $(1-\delta)$  values, averaged over 100 replicates with  $K_0 = 0.5$  and  $K_0 = 1.0$  for the separate sample dependent estimator using total population 15+ as the auxiliary variable for each of the Census Divisions (unplanned domains) in this study. These average  $(1-\delta)$  values indicate the degree of reliance of the sample dependent estimator on the synthetic component. As expected, domains consisting primarily of partial strata tend to place increased reliance on the synthetic component. Nevertheless, that reliance remains quite small. For example, with  $K_0 = 1$  the highest value it assumes is .28 for Census Division 218. -

Also as expected, the average  $(1-\delta)$  values for  $K_0 = 0.5$  are lower than those for  $K_0 = 1.0$ , implying the lower the value of  $K_0$  chosen, the lower would be the value of  $(1-\delta)$  and consequently less reliance (weight) on the synthetic component of the sample dependent estimator. However as illustrated in Table 1, a trade-off between bias and efficiency is involved since lower choices of  $K_0$  also result in reduced efficiency. The above values of  $K_0$  provide a reasonable degree of confidence for the type of domains discussed here. In general, however, other values of  $K_0$  may be chosen depending upon e.g. the size of the domain, sample size, strata sizes and their geographical configurations with respect to the domain.

#### 4.5 Concluding Remarks

1. The use of population by age/sex fares uniformly better than the other auxiliary variables, although gains over total population (15+) are mariginal.
2. The post-stratified domain estimator although more efficient as compared to the simple domain estimator, performs poorly as compared to the other three small area estimators investigated.
3. From the point of view of bias, the separate estimator has smaller relative bias as compared to the combined synthetic estimator. Further while average biases tend to be fairly small and tend to increase only slightly when the auxiliary information became out-of-date, biases for individual domains can be very high and change dramatically, frustrating efforts to identify 'outliers' where reduced reliance on synteic estimators should be made.
4. The combined composite estimator constructed as a linear combination of post-stratified and synthetic estimators is more efficient than either of its component estimators although only marginally so, as compared to the synthetic component, for optimum value of  $\alpha$ . Its bias would depend upon the weight attached to the synthetic component since the bias of the post-stratified estimator would generally be



negligible. Further, as the computation of the optimum  $\alpha$  is quite involved, in practice only an estimated value of  $\alpha$  may be used, resulting in a decrease in efficiency of this estimator.

5. The synthetic, composite and sample dependent estimator with  $K_0 = 1$  are all more or less equally efficient, and out-perform the unbiased design based estimator for planned domains.
6. Since the bias of the separate synthetic estimator is smaller than that of the combined synthetic estimator, the separate sample dependent estimator would result in smaller relative bias as compared to the combined sample dependent estimator. The bias of the separate to the combined sample dependent estimator. The bias of the separate post-stratified domain component can be controlled by collapsing those strata for which the intersection with the domain is very small. Thus considering all the three aspects, bias, mean square error and the computational complexities, the sample dependent estimator constructed at the stratum level using population by age and sex would seem to be a better choice.

## 5. FUTURE DIRECTION OF INVESTIGATION:

The study reported in this paper has focussed on evaluation of certain small area estimation methods using only census and survey data, in the context of the LFS, primarily for unplanned domains (type c). The estimators examined made use of synthetic and post-stratified domain estimators in different ways in an attempt to strike a balance between bias and mean square error. Below we point to directions which future investigations might take in efforts to develop statistically sound techniques for small area data in the Canadian context.

In the context of the Labour Force Survey, since the small area estimation methods for the unplanned domains have out-performed the unbiased design based estimates for comparable planned domains, it would be desirable to extend this investigation to certain small planned domains (type a) as well. In par-

ticular the sample dependent estimator considered here and other similar estimators discussed in the literature will be further investigated for the Labour Force characteristics. In addition these investigations should also be extended to other smaller surveys conducted by Statistics Canada for which small area data are in demand. Further work on development of methods of variance estimation to be used in practice for these estimators is also needed.

Other estimators which seem to be promising are the Structure Preserving Estimators (SPREE) suggested by Purcell and Kish (1980). In this approach the estimation process, specified by the association structure (i.e. the relationship between y and x variables at some previous time at domain level) and the allocation structure (i.e. the current relationship at the larger area level), preserves the earlier relationship present in the association structure without interfering with current information in the allocation structure. In the Canadian context, for characteristics for which large scale surveys (such as the Labour Force Survey) are undertaken regularly, it would seem the short term demand for data for domains of the size of FED's or Census Divisions may be met through the use of refined estimation techniques (and pooling of estimates over a period of time) utilizing census and survey data alone. However, for meeting such demands in the longer term and for other types of data based on smaller surveys and other types and sizes of domains, all three sources of data namely census, surveys and administrative files would have to be fully explored. Multi-variate linear regression estimators of the type considered by Erickson (1974) and Gonzalez and Hoza (1978) using data from all three sources should be studied in detail for their bias, mean square error and the computational complexities. Each of the three sources, with limitations of their own, when put together offer considerable potential for improvements in the sense that the weaknesses of one source can be the strengths of another. Hence there is reason for optimism that statistically sound techniques exploiting the strengths of data from different sources in an integrated fashion hold the future key to good quality small area data for a large variety of subject matters.

#### ACKNOWLEDGEMENT

Discussions with Mr. R. Platek have been beneficial in the finalization of this paper.



Table 1. Efficiencies of Small Area Estimators Relative to Direct Estimator  
- Nova Scotia Census Divisions (Auxiliary data up-to-date).

Characteristic	Auxiliary Variable	Level of Construction	Post-Stratified Domain	ESTIMATOR		
				Synthetic	Composite ( $\sigma^2 = 0.223$ )	Sample Dependent $K_0 = 0.5$ $K_0 = 1.0$
Employed	Dwelling	combined	4.58	10.17	10.92	9.17      10.42
	Population	"	4.92	10.75	10.58	10.50      11.67
	Population by age/sex	"	5.08	10.83	11.25	11.17      12.25
	Dwelling	separate	2.75	10.50	-	9.58      10.50
	Population	"	2.83	10.92	-	10.58      11.42
	Population by age/sex	"	2.83	11.00	-	11.00      11.75
Unemployed	Dwelling	combined	1.33	1.70	1.75	1.40      1.55
	Population	"	1.36	1.70	1.75	1.43      1.58
	Population by age/sex	"	1.36	1.70	1.75	1.43      1.58
	Dwelling	separate	1.30	1.69	-	1.48      1.58
	Population	"	1.33	1.69	-	1.51      1.61
	Population by age/sex	"	1.33	1.69	-	1.51      1.61

Table 2. Efficiencies of Small Area Estimators Relative to Direct Estimator  
- Nova Scotia Census Divisions (Auxiliary data out-of-date)

Characteristic	Auxiliary Variable	Level of Construction	ESTIMATOR		
			Post-Stratified Domain	Synthetic	Sample Dependent ( $K_0=1.0$ )
Employed	population	combined	3.47	4.73	4.07
	population by age/sex	"	3.60	4.73	4.23
Unemployed	population	combined	1.44	2.19	1.68
	population by age/sex	"	1.46	2.21	1.69

Table 3. % Average Relative Absolute Bias of Synthetic Estimators for FED's (using out-of-date, 15+ population as auxiliary information)

<u>Characteristic</u>	<u>Employed</u>		<u>Unemployed</u>	
	<u>Separate</u>	<u>Combined</u>	<u>Separate</u>	<u>Combined</u>
<u>Province</u>				
Newfoundland	1.20	2.19	1.75	3.17
Prince Edward Island	3.54	5.62	3.71	2.80
Nova Scotia	0.87	1.64	1.25	1.87
New Brunswick	2.95	2.54	6.35	7.04
Quebec	2.53	3.87	3.87	4.51
Ontario	2.17	3.52	3.01	4.25
Manitoba	1.42	2.28	2.22	3.44
Saskatchewan	2.41	2.65	4.35	4.85
Alberta	5.24	7.08	5.12	10.33
British Columbia	1.73	2.71	2.72	4.20

Table 4. FED's with Biases of Separate Synthetic Estimator  
for Unemployed Exceeding 10%

Category (i)			Category (ii)			Category (iii)		
% rel bias			% rel bias			% rel Bias		
FED	Up-to-date	Out-of-date	FED	Up-to-date	Out-of-date	FED	Up-to-date	Out-of-date
102	12.25	-3.90	414	1.57	17.78	301	12.71	25.85
104	-12.52	6.23	426	-7.38	-11.78	304	-11.29	-17.57
411	-13.10	-0.37	450	1.93	-11.35	412	-10.07	-15.80
436	10.48	-0.48	455	2.67	15.46	438	12.15	14.22
474	18.35	-6.04	460	-7.74	20.80	501	10.52	16.83
806	15.15	-0.20	504	7.19	41.05	579	15.50	10.59
			527	9.59	11.76	818	10.83	39.41
			605	3.63	14.29			
			701	4.90	17.74			
			804	0.78	15.85			
			813	-0.48	-15.46			

Category (i): bias exceeds 10% only when auxiliary information is up-to-date.  
(ii): bias exceeds 10% only when auxiliary information is out-of-date.  
(iii): bias exceeds 10% in both the cases.

Table 5. Average Reliance of Separate Sample Dependent Estimator on Synthetic Component: Nova Scotia Census Divisions.

Census Division	Reliance $(1-\delta)$ on Synthetic Component		Proportion of Census Division Population	
	$K_0=0.5$	$K_0=1.0$	Partial Strata	Complete Strata
201	.04	.15	1.00	-
202	.15	.20	.70	.30
203	.01	.12	1.00	-
204	.12	.22	1.00	-
205	.04	.14	1.00	-
206	.03	.08	.37	.63
207	.05	.07	.26	.74
210	.06	.09	.35	.65
211	.04	.05	.13	.87
212	.06	.10	.52	.48
213	.03	.16	1.00	-
214	.04	.16	1.00	-
215	.11	.21	1.00	-
216	.05	.15	1.00	-
217	.01	.01	.03	.97



## REFERENCES

- [1] Drew, J.D. and Choudhry, G.H. (1979), "Small Area Estimation", Technical Report, Census and Household Surveys Methods Division, Statistics Canada.
- [2] Ghangurde, P.D. and Singh, M.P. (1976), "Synthetic estimation in the LFS", Technical Report, Household Surveys Development Division, Statistics Canada.
- [3] Ghangurde, P.D. and Singh, M.P. (1977), "Synthetic Estimates in Periodic Household Surveys", Survey Methodology, Vol. 3, No. 1, 152-181, Statistics Canada.
- [4] Ghangurde, P.D. and Singh, M.P. (1978), "Evaluation of Efficiency of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 53-61.
- [5] Gonzalez, M.E. (1973), "Use and Evaluation of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 33-36.
- [6] Gonzalez, M.E. and Waksberg, J. (1973), "Estimation of the Error of Synthetic Estimates", paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.
- [7] Gonzalez, M.E. (1975), "Small Area Estimation of Unemployment", Proceedings of the American Statistical Association, Social Statistics Section, 437-460.
- [8] Gonzalez, M.E. and Hoza, C. (1978), "Small Area Estimation with Application to Unemployment and Housing Estimates", Journal of the American Statistical Association 73, 7-15.
- [9] Holt, T., Smith, T.M.F. and Tomberlin, T.J. (1979), "A Model Based Approach to Estimation for Small Sub-groups of a Population", Journal of

- [10] National Center for Health Statistics (1968), "Synthetic State Estimates of Disability", P.H.S. Publication No. 1759, U.S. Government Printing Office, Washington, D.C.
- [11] Platek, R. and Singh, M.P. (1976), "Methodology of the Canadian Labour Force Survey," Catalogue No. 71-526, Statistics Canada.
- [12] Purcell, N.J. and Linacre, S. (1976), "Techniques for the Estimation of Small Area Characteristics", paper presented at the 3rd Australian Statistical Conference, Melbourne, Australia.
- [13] Purcell, N.J. and Kish, L. (1979), "Estimation for Small Domains", Biometrics 35, 365-384.
- [14] Royall, R.M. (1973), "Discussion of two papers on Recent Developments in Estimation of Local Areas", Proceedings of the American Statistical Association, Social Statistics Section, 43-44.
- [15] Royall, R.M. (1978), "Prediction models in Small Area Estimation", NIDA Workshop on Synthetic Estimates, Princeton, N.J.
- [16] Sarndal, C.E. (1981), "When Robust Estimation is not an obvious answer: The case of the Synthetic Estimator versus Alternatives for Small Areas", Proceedings of the American Statistical Association, Survey Research Section.
- [17] Schaible, W.L., Brock, D.B. and Schnack, G.A. (1977), "An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics", Proceedings of the American Statistical Association, Social Statistics Section, 1017-1021.
- [18] Schaible, W.L. (1978), "Choosing Weights for Composite Estimators for Small Area Statistics", Proceedings of the American Statistical Association, Survey Research Section, 741-746.

- [19] Schaible, W.L. (1979), "A Composite Estimator for Small Area Statistics", in Synthetic Estimates for Small Areas (J. Steinberg, Ed.) National Institute on Drug Abuse Research Monograph No. 24, U.S. Government Printing Office, Washington, D.C., 36-53.
  
- [20] Singh, M.P. and Tessier, R. (1975), "Some Estimators for Domain Totals", Journal of The American Statistical Association 71, 322-325.
  
- [21] Sonquist, J.N. and Morgan J.A. (1964), "The Detection of Interaction Effects", Monograph no. 35, Survey Research Center, Institute for Social Research, University of Michigan.

## CHARACTERISTICS OF RESPONDENT AND NON-RESPONDENT HOUSEHOLDS IN THE CANADIAN LABOUR FORCE SURVEY

Elizabeth Clayton Paul and Murray Lawes<sup>1</sup>

This article presents findings from a study to characterize responding and non-responding households in the LFS. This study was motivated by two projects associated with the LFS Redesign, namely, the family estimation project and evaluation of non-response compensation procedures. However, the results of the study are of general interest in the assessment of the quality of data emanating from the LFS.

### 1. INTRODUCTION

Non-response is the lack of complete information for all selected units in a sample or census. The occurrence of non-response poses special problems for the producers and users of survey data. Non-response affects the quality of survey data in two basic ways. First, it reduces the effective sample size, resulting in loss of precision of the survey estimates. Second, to the extent that differences in the characteristics of respondent and non-respondent units are not properly accounted for in the estimation strategies, it may introduce a bias into the survey estimates. This paper focuses on the latter aspect of quality, specifically the characterization of respondent and non-respondent units in the Canadian Labour Force Survey (LFS). This information will provide some insight into the potential effect of non-response on the survey estimates and will suggest some variables which should be considered when compensating for non-response. Units were characterized by the variables size of household, economic family type, length of time in the survey, location, age of household members and labour force status of household members. This study is based on data derived from the LFS longitudinal data files. A statement of

---

<sup>1</sup> Elizabeth Clayton Paul, Economic Characteristics Staff, Statistics Canada and Murray Lawes, Census and Households Survey Methods Division, Statistics Canada.

major findings from this analysis is found in Section 2 followed by a brief description of the LFS, of the longitudinal files and the methodology used to characterize non-respondent households in Section 3. Section 4 then presents the derived data and resulting analysis. The final section briefly discusses the impact of the findings of this study on the quality of LFS data at the individual, family and household levels and suggests potential methods of dealing with non-response to alleviate or minimize deficiencies in the survey data arising due to non-response.

## 2. STATEMENT OF MAJOR FINDINGS

Within the LFS, non-response compensation procedures are based on the assumption that the characteristics of non-respondent households are similar to the characteristics of respondent households. Should this assumption prove incorrect, the non-response adjustment procedure will contribute to a bias in the survey estimates. It is impossible to determine the exact extent of this non-response bias. However, by examining longitudinal data on the survey life of a household, a profile of respondent and non-respondent households may be determined and the extent of differences evaluated.

Of the many variables examined in the characterization of respondent and non-respondent households, the variables month in sample, household size and labour force status of household members exhibited a definite trend in relation to response status. With respect to month in sample, the levels of non-response decreased as month in sample increased. Between months one and two the percentage of non-respondent households decreased sharply, and then gradually continued to decrease until month six, implying survey tenure is a critical factor in the determination of survey response. Thus any estimates by rotation number based on a non-response adjustment across all rotation groups may impart a slight bias to estimates on a rotation number basis.

Regarding household size, non-response decreased as household size increased. On a distributional basis there were almost twice as many households of size one for non-respondent households as for respondent households; and conversely, for households of size 5 and over, there were over twice as many



households for respondent households. The implication is that a non-response adjustment which does not take household size into consideration will, on average, represent non-respondent households by households which contain more household members than the non-respondent households.

The response patterns exhibited by household size and month in sample remained unchanged when the two characteristics were jointly examined. Since the analysis of these two variables, household size and month in sample, has shown a strong functional relationship with non-response, a non-response adjustment incorporating household size and month in sample should do much to alleviate discrepancies by rotation number in sample survey estimates of household and economic family units, and of characteristics dependent on these variables.

In addition to household size and month in sample, a relationship between non-response and labour force status was also exhibited, with particular reference to unemployment. For non-respondent households, the percentage of individuals classified as unemployed increased as month in sample increased, while the percentage for respondent households remained relatively stable. When the added dimension of household size was examined, a definite relationship was exhibited for households of size one with a slightly more variable pattern being exhibited for households of size two or more. For households of size one, the percentages of individuals classified as employed and unemployed were substantially greater for non-respondent households than for respondent. Also, the percentage of employed individuals decreased as month in sample increased; however, the percentage of unemployed increased. For households of size two or greater, the differences in the labour force distributions for respondent and non-respondent households were less pronounced than those for size one households, but the percentage of unemployed individuals in non-respondent households of size two or more did generally increase as month in sample increased.

Although there may be advantages in utilizing some variables relating to labour force activities in addition to household size and month in sample in the non-response adjustment process, and thus improving the labour force estimates; the desire for a general weight adjustment, the small sample size

at this level of aggregation, and the relatively low level of non-response currently experienced in the LFS may preclude the implementation of a non-response adjustment based on labour force status related variables. However, a non-response adjustment on the basis of household size and month in survey should have some benefits for the labour force estimates. Consequently, it may be feasible to consider adjustments for two groups of households, namely size one and size two or more, and for two survey tenures, namely one month and two months or more, in evaluating any improvements to the current LFS non-response adjustment process.

### 3. DATA SOURCE

#### 3.1 The Labour Force Survey

The LFS is a multi-stage stratified random sample with stratification occurring within the economic region level for each province. The final unit of sample selection is the dwelling. Each selected dwelling remains in the survey sample for six months. At the end of that time, these dwellings are replaced by another group of dwellings in such a manner that every month one-sixth of the sample is replaced or rotated. This implies that in any given month, there are six panels of dwellings in the LFS with each panel at various stages of aging. That is, one panel is in the survey for the first occasion (i.e., the birth rotation group), one panel for the second occasion,...., and one panel for the sixth occasion.

During one week each month, Survey Week<sup>1</sup>, LFS interviewers contact selected dwellings to obtain information on the composition, demographic variables and labour market activities of household members who are part of the survey universe<sup>2</sup>. For various reasons, interviewers are unable to obtain information from all selected dwellings. These dwellings where no interview is conducted are classified as vacant dwellings or non-respondent households<sup>3</sup>, depending on their occupancy status. For vacant dwellings, no response is obtainable or expected; whereas, for non-respondent households, survey information is missing. An adjustment<sup>4</sup> for non-response to compensate for this missing information is made at the data processing stage based on the assumption that

households which have been interviewed, i.e., respondent households, typify households which should have interviewed, i.e., non-respondent households. Should this assumption be false, then a bias is introduced into the survey estimates by this adjustment for non-response. This bias will increase as the rate of non-response increases. For this reason, it is important that the characteristics of non-respondent and respondent households be similar, and for this reason much effort is expended (successfully) in minimizing non-response.

### 3.2 Longitudinal Data File

Estimates based on monthly cross-sectional LFS data provide a static snapshot of the population and labour market for each month; however, by linking respondent information over the survey lifetime, a dynamic view of labour market activities is observed. In any given month, dwellings in one of the six rotation panels complete their six-month tenure in the survey. For dwellings in this panel, it is possible to trace the household composition and response pattern over the previous five months. This tracing is done by means of the Longitudinal Data File. The Longitudinal Data File is formed by concatenating the information on a given household over its six months of survey life.

In the LFS, dwellings and individuals are assigned unique identification codes. This affords a method of linking individual, household, and dwelling information over the six months a dwelling is in the survey, thus creating the Longitudinal Data File.

Initially, longitudinal records containing the six monthly response status codes are created for each dwelling. If a dwelling is respondent for one or more months, then individual records containing information on the household members who were living in the household at the time it was respondent are also included on the longitudinal file. However, if no response is indicated over the six months, only basic dwelling information is available for the dwelling. Thus, every individual who was a household member at some time over the six-month survey period is associated with a Longitudinal Data File



record. From this record, labour market activity and demographic information can be obtained for the months the individual was a responding household member. Based on this formulation of longitudinal data, examination of responding and non-responding households can occur and the characteristics of each response type evaluated.

### 3.3 Methodology for Deriving Estimates

In examining the characteristics of responding and non-responding households, the type of household response for each month was required. On a monthly basis, there are three types of dwelling responses: respondent, non-respondent, and vacant. Responding households are those where the LFS questionnaire is completed for all or some eligible household members. Non-respondent households are occupied by individuals who should be included in the survey but, for some reason, choose not to participate or are unable to participate due to existing circumstances. Vacant dwellings, on the other hand, are not occupied, or are occupied by individuals not included in the survey universe.

Thus, in determining the characteristics of responding and non-responding households, dwellings labelled as vacant were ignored.

To obtain the characteristics of responding households, the characteristics of individual household members who responded in the survey were examined; however, to obtain the characteristics of non-responding households, an imputation strategy was implemented. The characteristics of a non-responding household should be identical to or closely approximated by characteristics of individuals in that household in a month of response.

For those households who did respond at least once during the six months the household was in the survey, the months of response were the information donors for any months of non-response during the six months. In this manner, the characteristics of non-responding households were estimated. To impute for non-response by this method, it was imperative that a given household be respondent for at least one month; however, the household could have been

respondent for more than one month. If this latter situation occurred, the month of response closest to the month of non-response provided the donor information. If two months of response were equally close to a month of non-response, the month prior to the month of non-response was chosen as the donor month. The following algorithm summarizes this technique.

<u>Month of Non-Response</u>	<u>Ordering of months to check for donor information</u>
1	2, 3, 4, 5, 6
2	1, 3, 4, 5, 6
3	2, 4, 1, 5, 6
4	3, 5, 2, 6, 1
5	4, 6, 3, 2, 1
6	5, 4, 3, 2, 1

If there was no month of response available, then no imputation was performed and this household was excluded from this study.

### 3.4 Cautionary Note

If non-response rates based on this study are compared to non-response rates by rotation groups from the monthly LFS, they will differ in magnitude. The main source of difference is the exclusion of certain non-respondent households from this study of longitudinal data. As previously indicated, the ability to characterize a household in a month of non-response depended on the availability of respondent data in an alternative month for that household. That is, there had to be at least one month of response for a non-respondent household to be characterized. This implies that a household which was non-respondent, or a combination of non-respondent and vacant, for each of the six months it was part of the survey sample was excluded from this study. Thus, some non-respondent households which contributed to the monthly LFS measurement of non-response did not contribute to this longitudinal study of non-



response. Approximately 1.4% of the total sampled households were excluded on this basis.

Exclusion of some non-respondent households is the main reason for differences in data from this study and any other study on non-response which is based on the monthly LFS data. In addition to this source of discrepancy, the weighting technique applied may cause estimates to vary. For this report records were weighted by a product of the inverse sampling ratio, the sub-sampled cluster weight, and the stabilization weight<sup>5</sup>. In examining and interpreting the results in Section 4, or comparing these results to any other study on non-response, it is necessary to remember that the data source was the Longitudinal Data File, only records with at least one month of response contribute to the estimates, and the weighting structure was based on sample design weights only.

#### 4. ANALYSIS

The methodology in the previous section documented the procedures used to derive estimates of characteristic totals from the longitudinal file. In this section a number of variables (separately and jointly) are examined with respect to their characterizations between respondents and non-respondents. A particular variable or cross-classification of variables is dealt with in each of the following subsections. The motivation for examining the variables, tables containing relevant tabulations and a summary of the essential results are presented for the various subsections.

##### 4.1 Month in Sample

As noted in the introduction the LFS is based on a rotating panel design with each panel of dwellings remaining in the sample for a period of six months. At the sample design stage, considerable effort is taken to ensure that the sample associated with each rotation number (i.e. dwellings by panel) is a representative one-sixth subsample of the full LFS sample. In the past a number of references have been made to the phenomenon of rotation group bias, i.e. that the expected value of estimates based on a single panel differs

depending on number of months in the sample. For this reason the composition of the sample by month in sample and by response status were examined. Weighted estimates of the number of households at the Canada level by month in sample and by response status were obtained based on averages over 1980 and 1981 and are presented in Table 1. Due to design efforts to ensure representativeness of the sample by rotation number, it was expected that the total weighted counts would be equally distributed by month in sample. Examination of the data revealed that very close to one-sixth (or 16.67%) of the total households fall into each month in sample class. In all cases the differences in percentage distribution for a cell were within one-half of 1%.

When distributions of households by month in sample were examined by response status, deviations from a uniform distribution were observed, particularly for non-respondent units. The non-response rates by month in sample exemplified this fact. As illustrated in Table 1, the rate of non-response decreased as the number of months in the survey increased. The largest decrease occurred between the first and second months in the sample when the rate in the second month was approximately one-half of the rate in the first month. Further reductions in the non-response rates were observed as the number of months in the sample increased. Decreases in the rates between the second and sixth months were 21.1% and 34.2% for 1980 and 1981 respectively.

The percentage distribution of non-respondent households exhibited a similar decreasing trend as number of months in the sample increased. On a distributional basis, there are substantially more non-respondent households in the first month in sample than there were respondent households; however, this number decreased with increasing tenure in the survey. Thus any estimates by rotation number based on a non-response adjustment across all rotation numbers may impart a slight bias to estimates on a rotation number basis.

#### 4.2 Household Size

In the LFS, non-response generally occurs at the household level, i.e. the rate of partial non-response within households is very low. The household is the unit at which non-response occurs. Thus the characterization of house-

holds is necessary for the determination of the effects of non-response on estimates from the survey - be they at the level of household, family, or individual units. Perhaps the most basic household attribute, in relation to deriving demographic/socio-economic estimates from the survey, is household size. From a data collection point of view it is reasonable to assume that difficulties of contacting households decrease with increasing household size.

To evaluate the potential effect of household size on the non-response rate, Table 2 presents the percentage distribution of households by size and response status based on averages over the calendar years 1980 and 1981. For both years the non-response rate decreased dramatically as household size increased. Non-response rates by household size ranged from a high of 7.48% for households of size 1 to a low of 1.89% for households of size 5 or more in 1980 and correspondingly from 6.58% to 1.69% in 1981 for households of sizes 1 and 5 or more, respectively. An examination of the distribution of responding and non-responding households by size of household revealed a substantial difference in the distribution of households by size depending on the response status. On a distributional basis there were almost twice as many households of size one for non-respondent households as for respondent households. For respondent households there were slightly more than 50% which were of size 3 or more, whereas for non-respondent households only about 30% were of size 3 or more. The distributional differences in household size between respondent and non-respondent households was also reflected in the average household size for each response type. For 1980 the average household size for respondent and non-respondent households was 2.93 and 2.26, respectively; while for 1981 the corresponding sizes were 2.88 and 2.19. The implication is that with the adjustment for non-response at the LFS data processing stage, non-respondent households are represented by households which, on average, contain more household members than the non-respondent household. This leads one to question the assumption that respondent households typify non-respondent households, at least with respect to household size.



#### 4.3 Household Size by Month in Sample

In the previous two subsections substantial variations in the response rates were noted depending on the number of months in sample and also depending on the size of household. The next table was obtained to determine whether the noted variations in non-response rates were also observed when either household size or month in sample was held constant. Based on annual averages for 1980 and 1981, Table 3 presents percentage distributions of respondent and non-respondent households by household size and month in sample as well as the corresponding non-response rates for 1980 and 1981, respectively.

These tables show that the decreasing trends in non-response rates observed in Tables 1 and 2 for the full populations also hold true when the rates are examined holding one of the variables constant and letting the other vary. For example, in Table 1 non-response rates for all household sizes combined were shown to decrease as month in sample increased. Table 3 generally shows the same phenomenon when one examines the pattern of response rates by month in sample for each of the household size groupings separately. As when months in the survey alone were examined, the non-response rate decreased sharply from month one to month two. Similarly, the non-response rate decreased from month one to month two by approximately one half for each given household size. For households of size one and two the non-response rate continued to decrease in subsequent months in the survey; however, for households of size 3 and greater the non-response rate tended to stabilize during the second month in the survey.

Holding the number of months in the survey constant and examining the non-response rate as the household size varied, revealed a pattern similar to that exhibited in Table 2, where household size alone was considered. The non-response rate decreased with increasing household size. Table 3 likewise shows that for a given number of months in the survey (from one to six), there is a decreasing trend in the non-response rate as household size increases.

Combining these two trends, there was an expectation that the highest non-response rate would be observed in households of size one during the first

month in the survey. Similarly, there was an expectation that the lowest non-response rate would be observed in households with five or more members during the final month in the survey (i.e., in month six). Based on annual averages for 1980 and 1981, this expectation was verified. In 1980 and 1981 the non-response rates of highest magnitude were 13.39% and 12.81% respectively. Each of these rates applied to households of size one during the initial survey month. The non-response rate of least magnitude in 1980 was 1.54%. This applied to households containing five or more members during the third month in the survey; however, a non-response rate of 1.59% also applied to households containing five or more members for month 6. In 1981, the non-response rate of least magnitude was 1.37%. This occurred in households having five or more members during month 3, while the non-response rate for month 6 was 1.39%. Thus, although the lowest non-response rate did not uniquely occur in households containing five or more members during the final survey month, the non-response rate for households in this cell was not significantly different.

The distributions of household size by survey duration by response status indicated the potential for non-response bias in survey estimates. A non-response adjustment which does not take into account household size, will implicitly compensate for non-respondent households on the basis of the distribution of respondents, i.e., underestimating households of size 1 and 2 and over-estimating households of size 3 or more. It can be seen on a distributional basis that there were substantially more households of sizes 1 and 2 among non-responding households than there were among responding households and, of course, conversely fewer households of larger sizes (3, 4 and 5+) among the non-responding households than among the responding households. This discrepancy in distributions became more exaggerated when months in sample, or rotation groups, were considered, particularly for months one and two. After month two, the non-response rate tended to stabilize for households of size greater than two, whereas for households of size 1 or 2, the non-response rate continued to vary over the survey lifetime. This suggested that household size and rotation number are important characteristics to consider when methods for non-response adjustment are being evaluated.



#### 4.4 Family Composition of Household

In Section 4.2 there were substantial differences in the distribution of households by size between respondent and non-respondent households. To further evaluate household size discrepancies between respondent and non-respondent households, tabulations of households in terms of their composition of family types were obtained. The family type compositions were based on the number of economic families in the household, the size of the family units, the presence of children, and the marital status and age of the head of the family unit. The specific variables are indicated in Table 4a with corresponding percentage distributions and non-response rates by type by response status in Table 4b.

The higher non-response rates for households of size one were again evident from these tabulations. The rates were particularly high for households containing only an unattached individual aged less than 65 years of age. Households containing a married couple with other members present in the household (children or non-children) i.e., codes 6, 7 and 8 had low non-response rates relative to other types of households. In other words, there were proportionately more of these types of households among the responding than among the non-responding households. Households containing only unattached individuals (either one or more) and households containing a married couple only formed a higher percentage of non-responding households than of responding households. Thus in addition to household size, the composition of the household in terms of family types appeared to have some influence on the rate of non-response. Thus certain types of family units may not be properly compensated for in various weight adjustment strategies for non-response. This is particularly a crucial issue in the production of family estimates.

#### 4.5 Age of Individuals

Although the unit of potential response is generally the household, Table 5 presents percentage distributions by age group and response status at the individual level. Also presented are the distributions of the non-respondents

as percentages of the total population, or these could be referred to as individual level non-response rates.

The rate of non-response for all individuals combined were 3.13% and 2.63% for 1980 and 1981 respectively. These rates corresponded to household level non-response rates of 4.02% and 3.43% respectively for 1980 and 1981. The lower rates at the individual level were indicative of the inverse relationship between the size of household and the level of non-response as pointed out in Section 4.2. Since larger households had lower non-response rates, a greater proportion of individuals fell into the responding category. The relationships on a distributional basis between individual respondents and non-respondents bore out the results of the previous section with respect to the generally lower household non-response rates in households which contained children. For the age groups 0-14 and 15-19, the non-response rates in 1980 were 2.50% and 2.42% respectively, while in 1981 they were 2.12% and 1.92%. The highest non-response rates were observed in the age groups 65+ and 20-24. This again reflected the inverse relationship between household size and the non-response rate. Households of size 1 and 2 had the highest non-response rates. Individuals within the age groups 65+ and 20-24 were more likely to live alone or as a couple; hence, the non-response rates for these individuals were expected to be high. The variation in non-response rates by individual age groups indicates a potential effect on the quality of survey based estimates. In particular, age groups with a lower non-response rate than the over-all individual non-response rate will be over-estimated by a weight adjustment factor which does not take into account age variables. The opposite occurs when the non-response rate for the age group is greater than the overall individual non-response rate. To some extent any distortions introduced at the provincial level are corrected by the application of the ratio adjustment procedure.

#### 4.6 Age of Individuals by Size of Households

Continuing from the previous section the distributions of individuals by age groupings and response status were obtained within various household size breakdowns. These distributions as well as non-response rates, are presented

in Table 6a based on 1980 annual averages and Table 6b based on 1981 annual averages.

The distributions of individuals by age group were relatively similar by household size between respondents and non-respondents in households of sizes 2, 3, 4 and 5+; however, for households of size 1 there were substantial differences in the distributions. Within size 1 households the primary differences were for age group 25-44 in which there were substantially more individuals (on a distributional basis) in non-responding than responding households (39.6% compared with 28.8% for 1981 and 35.5% compared with 27.9% for 1980) and for age group 65+ in which there were substantially fewer individuals in non-responding households than in responding households (22.3% compared with 34.3% for 1981 and 22.4% compared with 34.3% for 1980). This latter observation is particularly important as about 28% of the population 65+ reside in households of size 1 whereas less than 5% of individuals in the age group 25-44 reside in households of size 1. Thus, it is differences in the distributions by age groups between respondents and non-respondents which merit special attention in any procedures to compensate for non-response in households of size 1.

The non-response rates in Tables 6a and 6b show that individual non-response rates within age groups exhibit the same pattern across household size measures as was observed in Section 4.2, namely that non-response rates decrease as household size increases. Within a particular size of household the relationships of non-response rates by age group were very different than non-response rates by age groups for all household sizes combined. Perhaps most notable was the fact that for each household size group separately (except size 4 in 1980), individuals 65+ exhibited the lowest level of non-response whereas the non-response rate for individuals 65+ in all households combined was the largest of any age group. This phenomenon resulted from the fact (mentioned earlier in this section) that the majority of individuals of age 65+ live in households of size 1 or 2, where the non-response rate was the greatest.



These tables indicate that non-response is very much dependent on household size and that age is not an important factor apart from the fact that there is a relationship between household size and the age of individuals residing in the household.

#### 4.7 Age of Individuals by Month in Survey

The distribution of individuals by age group for varying numbers of months in the survey, separately for respondents and non-respondents, are presented in Tables 7a and 7b for 1980 and 1981 respectively, as well as the corresponding non-response rates.

From Tables 7a and 7b it can be noted that distributions by age group for respondents were virtually identical regardless of the number of months in sample. Although the distributions for non-respondents showed a higher degree of variability for differing months in sample, there remained a degree of stability in the distributions. The pattern between distributions for respondents and for non-respondents was similar for each month in sample breakdown as it was for totals across months in sample.

A study of individual non-response rates again indicated in general a decreasing trend as number of months in sample increased. This occurred for individual age groups as well as for the total population. As expected the pattern over time was not as pronounced for individuals as it was on a household basis. This can be attributed to changes in the response pattern for various sized households; that is, there is a tendency for larger sized households to become non-respondents in the later survey months while smaller sized households tend to become respondent (refer to Table 3).

#### 4.8 Labour Force Status

In this subsection attention is turned from the basic demographic characteristics of households by response status to the characteristics of labour force activity. This evaluation was motivated by the desire to assess potential non-response bias in the survey estimates of these characteristics.

Section 4.2 presented substantial differences in the distributions of respondent and non-respondent households by household size, while Section 4.1 presented similar findings for month in sample. For this reason, the distributions of individuals by labour force status within each category defined by household size, month in sample, and response status were examined. They are presented in Table 8a.

Examination of these distributions by labour force status for all individuals regardless of size of household, showed that the distributions for respondent households differ in some important ways from the distributions of non-respondents and the pattern of differences was not consistent over time. The percentages of individuals unemployed showed perhaps the most interesting changes. For respondents, this percentage was relatively constant for each number of months in the sample; whereas, for non-respondent households, there was an increase in the percentage of individuals unemployed as the number of months in sample increased. The percentage of the population (aged 15 and over) unemployed for respondent households ranged from a low of 4.7% in months 3 to 6 to a high of 5.0% in month 1 for 1980, and a low of 4.6% in months 4 and 5 to a high of 4.9% in month 1 for 1981. For non-respondent households, the corresponding range of percentages was 4.5% in month 1 to 6.4% in month 6 for 1980, and a low of 4.0% in month 1 to a high of 6.2% in month 5 for 1981. A comparison of the percentage unemployed for each response status over time shows that there were fewer unemployed persons among non-respondent than respondent households for households in the sample for the first occasion and more unemployed persons among non-respondent than respondent households for households in the sample for four to six months. The relationship was variable for months two and three. A comparison of the percentage distribution patterns of labour force activities for respondent households over time indicated a relatively stationary distribution; however, the pattern for non-respondent households varied. For non-respondent households there were greater fluctuations in the percentage distributions for each labour force status across months. No distinct pattern of change was exhibited except with unemployment where representation increased with survey duration. This variation among non-respondents was at least partly attributable to small sample sizes of non-respondents relative to sample sizes for respondents.



Since unemployment is more sensitive to sample fluctuations than the other labour force statuses and exhibits a definite trend over time, compensating for non-response over rotation groups would distort this characteristic. Adjusting over rotation groups would result in an overestimation of unemployment in month 1, and an underestimation of unemployment in months 4 to 6. Since the divergence between responding and non-responding households in the percentage distribution of unemployment was more pronounced in the later survey months, the overall effect would be an underestimation of unemployment. Since the non-response adjustment occurs at the household level, not at the individual level, and the size of the household has proven to be an important response determinant (see Section 4.2), it is essential to consider household size as an additional component for the evaluation of non-response with respect to the labour force status.

When distributions by labour force status and month in sample were examined by household size breakdowns, the patterns or relationships noted above did not hold. For households of size 1, the proportions of individuals employed and unemployed were substantially higher for non-respondents than for respondents. For respondents the proportion of individuals employed and the proportion unemployed were relatively constant for varying number of months in the sample. For non-responding households, there was a general decrease in the proportion of individuals employed as the duration in sample increased; whereas, there was a substantial increase in the proportion of unemployed as the number of months in sample increased.

For households of other sizes (2,3,4 and 5+), the differences between labour force status distributions for respondent and non-respondent households were much smaller. Also patterns between distributions for respondents and non-respondents were not nearly as strong or consistent as for the case of household of size 1. On a distributional basis, there were generally fewer unemployed individuals in non-respondent households for the first survey occasion and more unemployed individuals in non-respondent households for the fourth and subsequent months in the sample, than for responding households. For households in the survey for two or three months the pattern was variable.

The percentage of individuals "not in the labour force" differed between responding and non-responding households by household size. In households of size 1 and 2 there were fewer individuals "not in the labour force" in non-responding households than in responding; whereas, no definite pattern existed for households of size 3 or more. As the employed constituted the majority of the group "in labour force", generally the relationship on a distributional basis between respondent and non-respondent households was the complement of that noted for the characteristic "not in the labour force".

Table 8b presents unemployment rates by household size and month in sample by response status for 1980 and 1981 respectively. These results are related to those in the previous tables and observations may be similar in that the relationship between unemployment rates for respondents and non-respondents are the result of the relationships between proportions employed and unemployed between respondent and non-respondent units.

For all individuals (i.e., regardless of household size) the rate of unemployment for non-respondents was less than the rate for respondents for the first month and greater than the rate for respondents in months 4 to 6. The relationship between the rates for months 2 and 3 varied by year. For non-responding households, there was a substantial increase in the unemployment rate as the number of months in the survey increased. This phenomenon was not observed for respondents where the first month in sample had the highest rate but the pattern for subsequent months was somewhat variable.

For households of sizes 2, 3, 4, and 5+ the same general relationship in unemployment rates between respondent and non-respondent households was observed as for the full set of individuals (i.e., regardless of household size). There was no definite pattern in unemployment rates over time for non-respondent households when various household sizes were considered. For households of size 1 the unemployment rate for non-respondents was generally higher than the rate for respondents.

#### 4.9 Type of Area

Results presented in Section 4.3 showed that there were substantial differences in distributions of households by size and month in sample between responding and non-responding households. This section further examines these results within broad types of area determined generally on the basis of population concentration and density; namely, self-representing areas (SRU), non-self representing urban areas (NSRU urban), and non-self-representing rural areas (NSRU rural). Although a more precise definition of area types is available, for this study it is sufficient to note that SRU's consist of the larger cities in the country, NSRU urban areas consist of smaller cities and towns, and NSRU rural areas are composed of the more sparsely populated portions of the country, including small villages and farm land. Due to the very small sample sizes, special areas were not considered. In very general terms, the patterns observed in Section 4.3 for all area types combined, were similar to those observed for the three broad area types; however, there were different distributions by household size for respondents depending on type of area. In SRU areas, on a distributional basis, most households were smaller sized whereas there were fewer smaller sized households in NSRU rural areas. The opposite was observed for larger sized households. The relationship between respondent and non-respondent households, however, was relatively the same regardless of type of area. From Tables 9a and 9b it can be noticed that there were approximately twice as many households of size 1 in non-responding households as in responding households and approximately one-half as many larger sized households (5+) in non-responding as in responding households.

Non-response rates, although levels differ by type of area, showed the same pattern of decreases by number of months in sample as was observed for all units combined (i.e., as compared with results presented in Section 4.3). Again there were substantial decreases in levels of non-response between the first and second months with decreases of lesser magnitude occurring in subsequent months.

The rates of non-response for all households (i.e., regardless of household size) were the highest for SRU areas, followed by NSRU urban areas and were



the lowest for NSRU rural areas. These differences were a function of the distributions of households by size across area types. Within specific size of household groupings, the patterns between respondent and non-respondent households are generally the same as when examined for comparable size groupings for all area types combined. The type of area variable is an important factor in compensation procedures as it differentiates between areas with different levels of non-response. However, in addition to size of households and month in sample variables the type of area variable does not provide much additional information in the characterization of survey units by response status.

## 5. SUMMARY

The previous section presented characterizations of responding and non-responding households with respect to a wide range of variables. The households and/or individuals displayed somewhat different characterizations depending on their response status. On the assumption that responding and non-responding households exhibit similar characteristics, it would seem to be important to incorporate some of the variables examined in Section 4 into non-response compensation procedures for the survey.

The method of compensating for non-respondent households in the LFS is carried out within small geographic areas (balancing units) by an inflation of the design weight by the inverse of the household response rate. These adjustments are made on the basis of household counts independent of any characteristics of the household. Unless there is a high degree of correlation among households within balancing units, one would expect very little reduction in non-response bias by the present adjustment procedure.

An indication of the magnitude of non-response bias under the current procedure for compensation for non-response would be desirable. An explicit imputation of missing information due to non-response on the LFS file can be obtained using procedures similar to those used in this study. After adjustments for complete non-response (i.e., non-response for all six months) survey estimates based on these comprehensive imputation strategies can be obtained.

Comparison of these resulting estimates with official survey estimates would provide added support to assessments of response bias which have been alluded to in this report.

This report has provided justifications for considering various additional variables in the adjustment for non-response: month in sample, household size and labour force status. As there are substantial variations in the response rate by rotation number (month in sample) it is advisable to adjust for non-response within each rotation number separately. As the pattern of labour force characteristics for non-respondents exhibits a degree of variation over months in sample, an adjustment on the basis of rotation number should have some benefits for labour force estimates as well. As the greatest differences are between the first month and subsequent months in sample, an adjustment for these two classes may be sufficient.

Among the non-responding households there are substantially more households of size one (and to a lesser extent for size two) than in responding households. Thus, household size is an important variable to be incorporated in any adjustment procedures for non-response. The analysis has shown that discrepancies are the greatest for households of size one. It may thus be feasible to consider adjustments for two groups of households only, namely households of size one and households of size two or more. Incorporation of household size into compensation procedures for household non-response necessitates having some information available about the size of non-responding households. This may be explicit, as for example the household size on a previous survey occasion, or implicit, as for example a distribution of non-responding households by size from previous surveys, or a distribution by household size from an independent source such as the Census. In either situation, adjustments incorporating considerations of household size in conjunction with adjustments by rotation number, should do much to alleviate discrepancies by rotation number in sample survey estimates of household and economic family units.

As noted in Subsection 4.9, even within household size and month in sample, there are differences in the distributions of respondents and non-respondents



by labour force status. For the LFS there may be advantages in utilizing some variables relating to labour force activities in the adjustment process. There are two factors which tend to preclude this as being viable in practice. Namely, there is a desire for a general weight adjustment, not only for the LFS but also for the various supplementary surveys, and secondly, information at this level of disaggregation would be very unstable and necessitate adjustments at higher levels of aggregation. This new level of adjustment would negate any advantages which may currently be experienced due to local labour market phenomenon. Any compensation procedures must bear in mind the relatively low level of non-response currently experienced for the LFS. This has implications on the level of sophistication warranted, the potential for impact on the estimates, and the reliability of non-response information which would form a key part of the procedure.

There are a range of possible alternatives to the present method of compensating for non-response. Further work in the development of other feasible compensation strategies is a two-staged process. The first stage is the simulation and evaluation of monthly labour force estimates based on the imputation strategy suggested in this report. The second stage is the development of other non-response adjustment strategies followed by their empirical evaluation. Such work is in fact under way.

#### FOOTNOTES

- [1] The estimates provided by the Labour Force Survey refer to the specific week covered by the survey each month, Reference Week, normally the week containing the 15th day. Survey Week, when all interviews are conducted, is the week immediately following Reference Week.
- [2] The survey universe for the Labour Force Survey is all persons in the population aged 15 years of age or over residing in Canada, with the exception of the following: residents of the Yukon and the Northwest Territories, persons living on Indian Reserves, inmates of institutions and full-time members of the Armed Forces.

- [3] Each month the interviewer is required to indicate whether a complete interview was obtained, that is, a complete Labour Force Survey questionnaire was completed for each eligible household member; a partial interview was obtained, that is a questionnaire was completed for some but not all eligible household members; or no interview was obtained. When no interview occurs, the interviewer must indicate the reason for this. Non-respondent households include those where no one was home (after several calls), the household refused to respond, the household was temporarily absent, or the interview was prevented by weather conditions, death, sickness, a language problem or other unusual circumstances in the household. Vacant dwellings include unoccupied dwellings, seasonal dwellings, dwellings under construction, dwellings occupied by persons not to be interviewed, and dwellings demolished, converted to business premises, moved, abandoned (unfit for habitation), or listed in error.
- [4] For further detail on the LFS non-response adjustment see "Methodology of the Canadian Labour Force Survey, (1976)", Statistics Canada, Catalogue 71-526 Occasional, October 1977, pp. 67-68.
- [5] For further detail on the LFS weighting process see "Methodology of the Canadian Labour Force Survey, (1976)", Statistics Canada, Catalogue 71-526 Occasional, October 1977, pp. 65-74.

TABLE 1. Percentage Distributions for Respondent and Non-respondent Households by Month in Sample for 1980 and 1981, Canada

Month in sample	Total	Respondent	Non-respondent	Non-response rate
<u>1980</u>				
1	16.6	16.1	28.6	6.94
2	16.6	16.7	15.9	3.84
3	16.7	16.8	14.4	3.47
4	16.7	16.8	14.3	3.45
5	16.7	16.8	14.2	3.42
6	16.8	16.9	12.6	3.03
Total	100	100	100	4.02
<u>1981</u>				
1	16.6	16.0	32.1	6.66
2	16.7	16.7	16.6	3.42
3	16.7	16.8	14.4	2.96
4	16.7	16.8	13.9	2.83
5	16.7	16.9	12.1	2.48
6	16.7	16.9	11.0	2.25
Total	100	100	100	3.43

TABLE 2. Percentage Distributions of Respondent and Non-respondent Households and Non-Response Rates by Household Size for 1980 and 1981 Annual Averages, Canada

	1980				1981			
	Total	Respondent	Non-respondent	Non-response rate	Total	Respondent	Non-respondent	Non-response rate
Size of household								
1	19.0	18.3	35.4	7.48	19.9	19.2	38.1	6.58
2	29.3	29.2	32.8	4.50	29.8	29.7	32.6	3.76
3	17.8	18.2	13.4	3.00	17.6	17.8	11.9	2.33
4	18.8	19.1	11.4	2.44	18.8	19.1	10.4	1.90
5+	14.9	15.2	7.0	1.89	14.0	14.2	6.9	1.69
Total	100	100	100	4.02	100	100	100	3.43
Average household size	2.91	2.93	2.26		2.86	2.88	2.19	

TABLE 3. Percentage Distributions of Respondent and Non-respondent Households by Household Size for Month in Sample for 1980 and 1981 Annual Averages, Canada

Month in sample	Household size					
	1980					
	1	2	3	4	5+	Total
<u>Respondent households</u>						
1	17.4	29.1	18.3	19.5	15.7	100.0
2	18.1	29.2	18.2	19.1	15.4	100.0
3	18.4	29.2	18.1	19.1	15.2	100.0
4	18.5	29.1	18.2	19.1	15.1	100.0
5	18.7	29.1	18.1	19.0	15.0	100.0
6	18.9	29.2	18.1	19.0	14.9	100.0
<u>Non-respondent households</u>						
1	36.1	32.3	13.2	11.2	7.2	100.0
2	38.9	32.7	12.1	9.6	6.7	100.0
3	35.9	32.2	13.5	11.8	6.6	100.0
4	34.9	33.2	13.4	11.7	6.9	100.0
5	32.9	33.9	13.8	12.5	6.9	100.0
6	32.1	33.0	14.9	12.4	7.7	100.0
<u>Non-response rates</u>						
1	13.39	7.64	5.10	4.10	3.30	6.94
2	7.90	4.28	2.59	1.97	1.71	3.84
3	6.56	3.81	2.61	2.17	1.54	3.47
4	6.32	3.92	2.56	2.15	1.61	3.45
5	5.87	3.97	2.63	2.28	1.61	3.42
6	5.05	3.41	2.51	2.00	1.59	3.03
Total	7.48	4.50	3.00	2.44	1.89	4.02
<u>1981</u>						
	1	2	3	4	5+	Total
<u>Respondent households</u>						
1	18.3	29.6	17.9	19.5	14.7	100.0
2	19.0	29.8	17.7	19.2	14.3	100.0
3	19.2	29.7	17.8	19.0	14.3	100.0
4	19.5	29.7	17.8	18.9	14.2	100.0
5	19.7	29.7	17.7	18.9	14.0	100.0
6	19.8	29.8	17.7	18.8	13.9	100.0
<u>Non-respondent households</u>						
1	37.7	32.9	12.8	10.3	6.3	100.0
2	40.5	32.7	11.5	9.1	6.2	100.0
3	41.5	33.0	9.6	9.4	6.5	100.0
4	37.6	31.7	12.5	11.0	7.1	100.0
5	36.5	32.6	12.0	10.9	8.0	100.0
6	34.1	32.4	12.3	12.7	8.5	100.0
<u>Non-response rates</u>						
1	12.81	7.34	4.85	3.63	2.97	6.66
2	7.03	3.74	2.25	1.65	1.51	3.42
3	6.19	3.28	1.62	1.49	1.37	2.96
4	5.32	3.02	2.01	1.67	1.44	2.83
5	4.50	2.72	1.70	1.45	1.43	2.48
6	3.82	2.45	1.58	1.53	1.39	2.25
Total	6.58	3.76	2.33	1.90	1.69	3.43



TABLE 4a. Determination of Family Type Composition Variable

Code	Number of economic family units in the household	Size of economic family unit	Age of head of family unit	Presence of children in the household	Head is a member of a married couple
1	1	1	25		
2	1	1	25-64		
3	1	1	65+		
4	1	2	45	No	Yes
5	1	2	45	No	Yes
6	1	2+	45	Yes	Yes
7	1	2+	45	Yes	Yes
8	1	2+		No	Yes
9	1	2+		No	No
10	1	2+		Yes	No
11	2+	all of size 1			
12	2+	all of size 2+			
13	2+	mixed			

TABLE 4b. Percentage Distribution of Respondent and Non-respondent Households by Economic Family Type for 1980 and 1981 Annual Average, Canada

Economic family type	1980			1981		
	Non-respondent households	Respondent households	Non-response rate	Non-respondent households	Respondent households	Non-response rate
	5.8	2.3	9.80	5.7	2.4	7.82
	21.0	9.5	8.51	23.4	9.9	7.71
	8.6	6.6	5.14	9.1	7.0	4.47
	9.5	8.0	4.72	9.5	8.0	4.05
	15.5	13.6	4.57	14.4	13.8	3.57
	18.4	28.1	2.67	16.3	27.0	2.10
	4.9	9.9	2.04	4.7	9.0	1.83
	4.6	8.2	2.32	4.1	8.6	1.65
	3.1	4.3	2.99	3.1	4.3	2.45
	3.9	4.8	3.30	5.0	4.9	3.47
	3.4	2.7	5.05	3.5	2.8	4.25
	0.0	0.1	1.25	0.1	0.1	2.29
	1.2	2.1	2.47	1.2	2.1	2.06
Total	100.0	100.0		100.0	100.0	

TABLE 5. Percentage Distributions of Individuals by Age Groups by Response Status for 1980 and 1981 Annual Averages, Canada

Age group	1980			1981		
	Respondent	Non-respondent	Non-response rate	Respondent	Non-respondent	Non-response rate
0-14	24.3	19.4	2.50	23.6	18.9	2.12
15-19	9.7	7.5	2.42	9.5	6.9	1.92
20-24	9.1	10.4	3.54	9.4	10.9	3.04
25-44	29.2	31.6	3.38	29.6	33.0	2.91
45-64	18.9	20.9	3.45	19.1	19.9	2.74
65+	8.7	10.2	3.65	8.9	10.5	3.08
Total	100.0	100.0	3.13	100.0	100.0	2.63

TABLE 6a. Percentage Distribution of Individuals by Age Group and Non-response Rates for Household Size and Response Status for 1980 Annual Averages, Canada

Age group	Household size					Total
	1	2	3	4	5+	
<u>Respondent</u>						
0-14	0.0	2.6	20.5	35.1	37.1	24.3
15-19	1.9	3.5	8.2	9.7	16.7	9.7
20-24	10.5	14.2	11.4	6.0	6.7	9.1
25-44	27.9	25.3	31.3	35.4	25.2	29.2
45-64	25.4	31.6	23.5	12.4	11.8	18.9
65+	34.3	22.8	5.2	1.3	2.6	8.7
Total	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-respondent</u>						
0-14	0.3	2.6	22.6	37.0	40.8	19.4
15-19	3.0	3.8	8.1	8.8	15.7	7.5
20-24	13.6	14.4	12.0	4.8	5.7	10.4
25-44	35.5	27.3	33.5	37.4	26.6	31.6
45-64	25.2	33.0	19.6	10.7	10.2	20.9
65+	22.4	19.1	4.1	1.4	1.1	10.2
Total	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-response rates</u>						
0-14	-	4.52	3.30	2.57	2.05	2.50
15-19	11.07	4.83	2.93	2.22	1.76	2.42
20-24	9.49	7.94	3.16	1.95	1.60	3.54
25-44	9.33	4.83	3.21	2.58	1.98	3.38
45-64	7.45	4.68	2.52	2.11	1.62	3.45
65+	5.01	3.80	2.41	2.47	0.85	3.65
Total	7.48	4.50	3.00	2.44	1.87	3.13

TABLE 6b. Percentage Distribution of Individuals and Non-response Rates by Age Group for Household Size and Response Status for 1981 Annual Averages, Canada

Age group	Household size					Total
	1	2	3	4	5+	
<u>Respondent</u>						
0-14	0.0	2.7	19.8	34.4	36.9	23.6
15-19	1.8	3.4	8.6	9.9	16.1	9.5
20-24	10.6	14.1	11.4	6.4	7.1	9.4
25-44	28.8	25.9	31.8	35.2	25.8	29.6
45-64	24.4	31.7	23.5	12.6	11.6	19.1
65+	34.3	22.2	5.0	1.5	2.5	8.9
Total	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-respondent</u>						
0-14	0.1	3.8	23.6	37.2	39.4	18.9
15-19	2.0	3.1	7.5	8.8	15.6	6.9
20-24	13.0	15.1	12.0	5.9	5.8	10.9
25-44	39.6	28.6	34.3	37.2	27.7	33.0
45-64	22.9	30.1	19.8	10.1	10.2	19.9
65+	22.3	19.3	2.8	0.9	1.2	10.5
Total	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-response rates</u>						
0-14	-	5.16	2.77	2.05	1.77	2.12
15-19	7.15	3.47	2.03	1.69	1.61	1.92
20-24	7.94	4.02	2.47	1.77	1.37	3.04
25-44	8.85	4.14	2.51	2.01	1.78	2.91
45-64	6.20	3.57	1.97	1.53	1.46	2.74
65+	4.38	3.27	1.31	1.15	0.83	3.08
Total	6.58	3.76	2.33	1.90	1.66	2.63

TABLE 7a. Percentage Distribution of Individuals and Non-response Rates by Age Group for Month in Sample and Response Status for 1980 Annual Averages, Canada

Age group	Month in sample						Total
	1	2	3	4	5	6	
<u>Respondent</u>							
0-14	24.2	24.1	24.3	24.4	24.5	24.5	24.3
15-19	9.9	9.8	9.7	9.7	9.7	9.6	9.7
20-24	9.2	9.2	9.2	9.1	9.1	9.0	9.1
25-44	29.1	29.1	29.2	29.3	29.2	29.1	29.2
45-64	18.9	18.9	18.9	18.9	18.9	19.0	18.9
65+	8.7	8.8	8.7	8.7	8.7	8.7	8.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-respondent</u>							
0-14	18.9	19.0	19.6	19.5	20.0	19.9	19.4
15-19	7.6	6.3	7.8	7.7	7.6	7.9	7.5
20-24	10.6	10.3	10.1	10.3	10.5	10.3	10.4
25-44	31.9	33.1	30.8	30.6	31.4	31.5	31.6
45-64	20.4	20.6	21.8	21.5	20.9	21.0	20.9
65+	10.6	10.8	10.0	10.4	9.6	9.5	10.2
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-response rates</u>							
0-14	4.22	2.25	2.18	2.16	2.24	2.00	2.50
15-19	4.12	1.84	2.15	2.16	2.17	2.03	2.42
20-24	6.12	3.16	2.94	3.04	3.16	2.78	3.54
25-44	5.83	3.21	2.83	2.81	2.93	2.65	3.38
45-64	5.74	3.08	3.09	3.06	3.00	2.71	3.45
65+	6.40	3.50	3.06	3.22	3.01	2.69	3.65
Total	5.34	2.84	2.69	2.70	2.73	2.46	3.13



TABLE 7b. Percentage Distribution of Individuals and Non-response Rates by Age Group for Month in Sample and Response Status for 1981 Annual Averages, Canada

Age group	Month in sample						Total
	1	2	3	4	5	6	
<u>Respondent</u>							
0-14	23.4	23.4	23.5	23.6	23.7	23.8	23.6
15-19	9.7	9.6	9.5	9.5	9.4	9.3	9.5
20-24	9.4	9.4	9.4	9.4	9.4	9.3	9.4
25-44	29.5	29.7	29.7	29.6	29.6	29.6	29.6
45-64	19.1	19.1	19.0	19.1	19.1	19.1	19.1
65+	8.9	8.9	8.9	8.9	8.9	8.9	8.9
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-respondent</u>							
0-14	18.1	18.7	18.3	19.8	19.8	20.1	18.9
15-19	6.9	6.2	6.8	6.6	7.3	7.6	6.9
20-24	10.9	10.5	11.1	11.0	11.2	10.7	10.9
25-44	33.5	33.2	32.0	32.6	33.1	32.6	33.0
45-64	20.2	20.5	20.8	19.2	18.9	19.3	19.9
65+	10.4	11.0	11.0	10.8	9.8	9.7	10.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
<u>Non-response rates</u>							
0-14	3.96	2.03	1.71	1.85	1.64	1.57	2.12
15-19	3.67	1.66	1.57	1.55	1.54	1.50	1.92
20-24	5.83	2.80	2.55	2.58	2.34	2.12	3.04
25-44	5.67	2.82	2.35	2.42	2.20	2.03	2.91
45-64	5.32	2.71	2.38	2.21	1.94	1.86	2.74
65+	5.88	3.10	2.68	2.66	2.15	2.01	3.08
Total	5.05	2.53	2.18	2.20	1.96	1.85	2.63

TABLE 8a. Percentage Distribution of Individuals by Labour Force Status for Month in Sample, Household Size and Response Status for 1980 and 1981  
Annual Average, Canada

Month in Sample																			
Size of Household	Response Status	1			2			3			4			5			6		
		E	U	N	E	U	N	E	U	N	E	U	N	E	U	N	E	U	N
<u>1980</u>																			
1	Respondent	51.0	3.2	45.8	52.3	3.2	44.5	52.2	3.2	44.6	52.2	3.2	44.5	52.4	3.3	44.5	52.2	3.2	44.6
	Non-respondent	65.0	4.4	30.6	62.8	4.9	32.3	60.2	6.8	33.0	60.0	6.3	33.8	60.9	6.2	32.9	57.0	8.1	34.9
2	Respondent	55.5	4.1	40.3	55.7	4.2	40.2	56.1	3.9	40.1	56.0	3.9	40.0	55.6	4.0	40.4	55.5	4.0	40.5
	Non-respondent	58.0	4.0	38.0	56.0	4.1	39.8	56.3	4.6	39.1	55.2	5.3	39.5	58.4	5.1	36.5	60.3	4.9	34.8
3	Respondent	61.7	5.5	32.9	61.9	5.2	32.9	61.9	5.2	33.0	61.9	5.2	32.9	61.6	5.4	33.0	61.9	5.2	33.0
	Non-respondent	62.4	4.5	33.1	57.9	6.7	35.4	59.5	5.0	35.5	58.4	5.3	36.3	61.6	5.9	32.4	58.7	6.3	35.0
4	Respondent	64.1	5.2	30.7	64.2	5.0	30.8	65.0	4.6	30.5	64.8	4.8	30.5	65.1	4.7	30.2	65.1	4.6	30.3
	Non-respondent	64.6	4.6	30.8	64.1	5.1	30.7	57.8	6.5	35.7	63.3	4.9	31.8	62.0	6.2	31.8	61.7	8.1	30.2
5+	Respondent	58.0	5.9	36.1	58.0	5.8	36.2	58.1	5.7	36.3	58.4	5.5	36.2	58.6	5.5	36.0	58.4	5.7	35.9
	Non-respondent	56.8	5.6	37.6	57.9	4.7	37.4	57.1	4.0	39.0	58.4	5.9	35.7	56.6	8.4	35.1	60.1	6.2	33.7
Total	Respondent	58.9	5.0	36.1	59.1	4.9	36.0	59.3	4.7	36.0	59.4	4.7	36.0	59.3	4.7	36.0	59.3	4.7	36.0
	Non-respondent	61.0	4.5	34.6	59.2	4.9	36.0	58.0	5.3	36.7	58.4	5.5	36.1	59.8	6.0	34.2	59.7	6.4	33.9
<u>1981</u>																			
1	Respondent	50.5	3.6	46.0	51.9	3.2	44.8	52.1	3.5	44.5	52.2	3.3	44.6	51.9	3.4	44.7	51.7	3.5	44.8
	Non-respondent	63.7	3.7	32.5	63.4	4.1	32.5	62.3	3.5	34.2	60.8	6.4	32.8	59.0	4.8	36.2	60.8	5.6	33.7
2	Respondent	55.7	3.9	40.4	56.1	3.8	40.1	56.1	3.7	40.3	56.0	3.7	40.3	55.8	3.8	40.4	55.7	3.9	40.4
	Non-respondent	60.4	3.5	36.1	56.9	4.2	38.9	58.4	4.5	37.0	56.4	4.7	38.9	59.2	5.9	35.0	56.5	5.0	38.5
3	Respondent	63.3	5.3	31.4	63.4	5.2	31.4	63.9	5.1	31.1	63.6	5.0	31.4	63.3	5.2	31.5	63.2	5.2	31.6
	Non-respondent	66.6	4.8	28.6	65.5	5.3	29.2	63.8	6.5	29.8	65.9	5.8	28.3	61.5	7.1	31.4	66.1	6.7	27.2
4	Respondent	64.8	5.1	30.1	65.1	4.9	30.0	65.1	5.1	29.8	65.4	4.9	29.6	65.5	4.8	29.6	65.6	5.0	29.4
	Non-respondent	63.4	4.2	32.4	66.4	4.3	29.3	61.6	3.9	34.6	63.3	5.6	31.2	62.3	6.8	30.9	63.2	6.9	30.0
5+	Respondent	59.5	6.1	34.4	59.6	5.8	34.6	59.8	5.8	34.5	60.2	5.6	34.2	60.4	5.4	34.2	60.2	5.9	34.0
	Non-respondent	62.3	4.8	32.9	60.1	5.6	34.3	60.2	5.4	34.3	58.5	5.3	36.2	61.6	7.5	31.0	61.8	6.7	31.5
Total	Respondent	59.7	4.9	35.3	60.0	4.8	35.3	60.1	4.7	35.1	60.2	4.6	35.2	60.1	4.6	35.2	60.0	4.8	35.2
	Non-respondent	62.8	4.0	33.2	61.3	4.5	34.2	60.7	4.5	34.8	60.2	5.5	34.4	60.3	6.2	33.5	60.7	5.9	33.4

TABLE 8b. Unemployment Rates by Household Size, for Month in Sample and Response Status for 1980 and 1981 Annual Averages, Canada

Household size	Response status	Month in sample					
		1	2	3	4	5	6
<u>1980</u>							
1	Respondent	5.88	5.74	5.77	5.83	5.99	5.79
	Non-respondent	6.36	7.16	10.16	9.44	9.29	12.45
2	Respondent	6.94	6.95	6.46	6.54	7.01	6.75
	Non-respondent	6.39	6.81	7.58	8.77	7.97	7.56
3	Respondent	8.14	7.76	7.71	7.71	8.04	7.69
	Non-respondent	6.75	10.37	7.74	8.27	8.79	9.65
4	Respondent	7.54	7.24	6.55	6.84	6.71	6.60
	Non-respondent	6.71	7.42	10.09	7.22	9.05	11.58
5+	Respondent	9.21	9.10	8.88	8.54	8.57	8.92
	Non-respondent	8.92	7.49	6.50	9.19	12.85	9.36
Total	Respondent	7.83	7.63	7.27	7.28	7.37	7.35
	Non-respondent	6.82	7.63	8.43	8.60	9.13	9.72
<u>1981</u>							
1	Respondent	6.56	5.86	6.22	5.87	6.16	6.38
	Non-respondent	5.52	6.07	5.29	9.46	7.53	8.39
2	Respondent	6.57	6.38	6.11	6.13	6.37	6.53
	Non-respondent	5.51	6.93	7.18	7.69	9.06	8.09
3	Respondent	7.77	7.59	7.35	7.31	7.57	7.60
	Non-respondent	6.76	7.46	9.19	8.12	10.33	9.25
4	Respondent	7.30	7.02	7.28	7.02	6.87	7.12
	Non-respondent	6.28	6.05	5.88	8.07	9.77	9.80
5+	Respondent	9.24	8.90	8.78	8.57	8.26	8.89
	Non-respondent	7.16	8.58	8.27	8.35	10.84	9.76
Total	Respondent	7.64	7.33	7.28	7.14	7.16	7.42
	Non-respondent	6.05	6.89	6.95	8.30	9.36	8.91



[illegible]



TABLE 9c. Household Non-response Rates by Type of Area, Household Size, and Month in Sample for 1980 and 1981 Annual Averages, Canada

Month in sample	Household size	Type of area					
		1980			1981		
		SRU	NSRU urban	NSRU rural	SRU	NSRU urban	NSRU rural
1	1	13.99	11.02	11.62	13.57	11.79	9.37
	2	7.82	7.31	7.05	7.98	6.84	5.90
	3	4.98	6.05	5.03	4.93	5.62	5.02
	4	4.03	4.48	4.15	3.59	3.76	3.45
	5+	3.13	4.01	3.45	3.21	3.10	2.61
	Total	5.32	6.65	5.79	7.25	6.42	4.95
2	1	8.21	6.79	7.20	6.30	6.32	6.56
	2	4.30	4.41	4.13	3.96	3.77	3.47
	3	2.44	3.12	2.82	2.24	2.24	2.69
	4	1.90	2.48	1.96	1.41	2.56	2.31
	5+	1.63	2.23	1.71	1.53	1.97	1.36
	Total	3.98	3.93	3.26	3.59	3.50	3.01
3	1	6.68	6.33	6.50	6.33	5.61	5.52
	2	3.83	3.67	3.88	3.44	3.20	2.81
	3	2.48	3.18	2.78	1.58	1.32	2.10
	4	2.08	2.60	2.27	1.46	1.21	1.93
	5+	1.35	1.81	1.87	1.35	2.16	1.37
	Total	3.53	3.60	3.20	3.13	2.81	2.52
4	1	6.41	6.58	6.04	5.50	5.26	5.02
	2	4.00	3.63	3.77	3.06	3.11	2.90
	3	2.44	2.71	2.91	2.05	1.99	1.87
	4	2.05	2.82	2.07	1.74	1.66	1.79
	5+	1.47	1.85	1.90	1.51	1.59	1.23
	Total	3.53	3.61	3.11	2.99	2.83	2.39
5	1	6.04	5.94	5.51	4.59	4.47	4.39
	2	4.05	3.81	3.81	2.79	3.15	2.55
	3	2.51	2.79	2.99	1.60	2.18	1.86
	4	2.22	2.56	2.34	1.40	2.10	1.52
	5+	1.41	1.75	1.94	1.64	1.17	1.12
	Total	3.51	3.50	3.14	2.59	2.75	2.14
6	1	5.12	5.54	5.30	3.92	3.48	4.24
	2	3.43	4.08	3.25	2.42	3.06	2.60
	3	2.45	2.51	2.95	1.58	1.48	1.63
	4	1.99	1.93	2.07	1.52	1.58	1.86
	5+	1.61	1.28	1.68	1.55	1.19	1.21
	Total	3.11	3.26	2.85	2.34	2.32	2.19

ROTATION GROUP BIAS IN THE LFS ESTIMATES<sup>1</sup>P.D. GHANGURDE<sup>2</sup>

The paper attempts to evaluate the impact of non-response adjustment by rotation groups on rotation group bias in the estimates from the Canadian Labour Force Survey. Results on bias and non-response characteristics are presented and discussed. An index used to measure rotation group bias is given and some empirical results are analyzed.

## 1. INTRODUCTION

In the Canadian Labour Force Survey (LFS) sample design each month one-sixth of the households rotate out of the sample and one-sixth rotate in. The sample is thus composed of six panels or rotation groups. In any given month households in a rotation group have been in the survey from one to six months, including the current month. It is well-known that in household surveys with rotation sample designs estimates for the same characteristics from different rotation groups could have different expected values. This phenomenon, called rotation group bias, has been studied for the LFS and other household surveys with rotation sample designs (see [1], [5], [7] and [8]).

Rotation group bias can be attributed to several factors. In the LFS the non-response rates at household level are known to differ between rotation groups i.e. number of months a household is in the survey. It is also known that non-respondent households tend to have different characteristics as compared to respondent households. Both these factors can contribute to bias. Due to conditioning of the respondent or familiarity with the survey

---

<sup>1</sup> Presented at the American Statistical Association Meeting in Cincinnati, August 1982.

<sup>2</sup> P.D. Ghangurde, Census and Household Survey Methods Division, Statistics Canada.

over a period of six months, response bias in the data from successive months can be of different magnitude. There is some evidence from the LFS reinterview data of such differential bias over the period of six months. However, in the literature it has also been hypothesized that rotation group biases can be attributed to differences in non-response probabilities between rotation groups [7]. Although individual probabilities are not known, their averages can be estimated by non-response rates.

In this paper an attempt is made to evaluate the impact on rotation group bias of non-response adjustment by rotation groups. In section 2 some results on bias are introduced and their implications on the bias in the estimates from different rotation groups are discussed. Section 3 presents some data on nonresponse rates in the LFS and characteristics of respondents and non-respondents by months in the survey and their contribution to rotation group bias. Section 4 explains the adjustment of LFS weight for non-response by rotation groups and its impact on the rotation group bias and an index used as a measure of rotation group bias. In section 5, some data on the index for labour force status categories, based on 1981 surveys, are analyzed.

## 2. THE STATISTICAL MODEL

We introduce a model which provides expressions for contribution to bias of differences in non-response rates, differences in characteristics of respondents and non-respondents and response bias for any groups of the sample in which adjustment of weight for non-response can be done. Rotation groups can be considered as a particular case of these groups.

A population of size  $N$  is assumed to be divided into "strata" of respondents and non-respondents of sizes  $N_1$  and  $N_2$  respectively. A simple random sample of size  $n$  is drawn and responses are obtained from  $n_1$  units and  $(n-n_1)$  units are non-respondents.

Suppose the sample can be divided into  $K$  groups such that non-response rates and characteristics of respondents and non-respondents differ between the groups. The data collection methods used in these groups and the extent of conditioning of respondents or their familiarity with the survey could be

different leading to differences in non-response rates and characteristics and also possibly to different response biases. By an extension of a result in [2] and [6] to include response bias component, the bias of the sample mean  $\bar{y}$  of  $n_1$  units (without adjustment of weight for non-response within groups) is given by

$$B(\bar{y}) = \frac{1}{\bar{R}} \sum_{i=1}^K P_i \bar{Y}_{1i} (R_i - \bar{R}) + \sum_{i=1}^K P_i (1-R_i) (\bar{Y}_{1i} - \bar{Y}_{2i}) \\ + \frac{1}{\bar{R}} \sum_{i=1}^K \bar{P}_i R_i \beta_i, \quad (1)$$

where  $\bar{Y}_{1i}$  and  $\bar{Y}_{2i}$  are population means of respondents and non-respondents in the  $i^{\text{th}}$  group,  $R_i$ , response rate for the  $i^{\text{th}}$  group,  $P_i$ , proportion of total population in the  $i^{\text{th}}$  group,  $\bar{\beta}_i$  mean response bias in the  $i^{\text{th}}$  group and  $\bar{R} = \sum_{i=1}^K P_i R_i$ , overall response rate.

The above expression shows the decomposition of bias into three components. The first shows contribution of differential response rates, the second due to differences in characteristics between respondents and non-respondents and the third due to response bias. For simplicity, we consider in this paper characteristics based on attributes, e.g., proportions of "employed" and "unemployed". We now consider the estimate  $\bar{y}_a$ , with adjustment for non response by inverse of response rate done within each group. Thus

$$\bar{y}_a = \frac{1}{n} \sum_{i=1}^K n_{.i} \bar{y}_i,$$

where  $n_{.i}$  is sample size in the  $i^{\text{th}}$  group and  $\bar{y}_i$  is mean of  $n_{1i}$  units in the  $i^{\text{th}}$  group. The bias of  $\bar{y}_a$  is given by

$$B(\bar{y}_a) = \sum_{i=1}^K P_i (1-R_i) (\bar{Y}_{1i} - \bar{Y}_{2i}) + \sum_{i=1}^K P_i \bar{\beta}_i. \quad (2)$$



The first component of bias in (1) due to differential response rates between groups is eliminated, the second component due to differences in characteristics remains the same and the third component due to response bias could be different from that in (1).

Based on a framework of response non-response error model involving response probabilities at unit level, the bias has been decomposed into components due to non-response and response errors [3]. The above decomposition of bias does not use response probabilities at the level of individual units but is simple enough for empirical evaluation of the components.

If response rates do not differ between the groups the first component is zero so that, (1) is identical to (2); hence non-response adjustment within the groups does not lead to reduction in bias. The difference in the bias of  $\bar{y}$  and  $\bar{y}_a$  is given by

$$B(\bar{y}) - B(\bar{y}_a) = \frac{1}{\bar{R}} \sum_{i=1}^K P_i (\bar{R}_i - \bar{R}) (\bar{Y}_{1i} + \bar{\beta}_i). \quad (3)$$

Thus if response rates are different, and  $\bar{Y}_{1i}$  and  $\bar{\beta}_i$  do not differ between the groups, there is no change in the bias after non-response adjustment within the groups. If the means  $\bar{Y}_{1i}$  and  $\bar{\beta}_i$  differ between the groups there is a decrease in bias if the term on the right-hand side of (3) is positive and an increase, if it is negative. The change in absolute bias from  $|B(\bar{y})|$  to  $|B(\bar{y}_a)|$  as result of adjustment will depend upon the sign and magnitude of the term on the right hand side of (3).

The bias of estimate of mean for  $i$ th rotation group, without adjustment and with adjustment of weight for non-response by rotation groups, is obtained from (1) and (2) by simple substitution of  $P_i = 1$  and keeping the terms corresponding to the rotation group. Also, from (3) the difference in biases of estimate for  $i$ th rotation group is given by



$$B(\bar{y}_i) - B(\bar{y}_{ia}) = \left( \frac{R_i - \bar{R}}{\bar{R}} \right) (\bar{Y}_{1i} + \bar{\beta}_i), \quad (4)$$

where  $\bar{y}_i$  and  $\bar{y}_{ia}$  are estimates for  $i^{\text{th}}$  rotation group before and after adjustment. Assuming  $(\bar{Y}_{1i} + \bar{\beta}_i) > 0$  for all  $i$ , if  $R_i < \bar{R}$ , the bias for  $i^{\text{th}}$  rotation group increases after adjustment and if  $R_i > \bar{R}$ , it decreases.

Since the population of respondents in a survey month is the same for various rotation groups, it may be argued that the proportions  $\bar{Y}_{1i}$  could be the same for all rotation groups or months in the survey. However, the differences in exposure to survey or conditioning of the respondents can produce different response biases,  $\bar{\beta}_i$ , between rotation groups. Thus the difference in the bias of  $\bar{y}$  and  $\bar{y}_a$  is given by

$$B(\bar{y}) - B(\bar{y}_a) = \frac{1}{\bar{R}} \sum_{i=1}^K P_i (R_i - \bar{R}) \bar{\beta}_i. \quad (5)$$

However, the difference in bias of estimates for rotation group  $i$  is given by (4).

It may also be noted that under the assumption of constant  $\bar{Y}_{1i}$  and  $\bar{\beta}_i$  for all  $i$  and differential response rates, non-response adjustment by rotation groups does not change the bias of estimate based on all rotation groups. However, the change in the biases of individual rotation groups after non-response adjustment are accounted for by different response rates.

The above results are useful in the evaluation of contribution of various factors to rotation group bias and the impact of adjustment of weight by rotation groups on the estimates of rotation group bias.

The LFS is a monthly national household survey with a sample size of 55,000 households. Each of the ten provinces in Canada is divided into economic regions, which consist of groups of counties with similar economic structure. The economic regions are divided into homogeneous strata on the basis of distribution of employed persons in various industry-occupation groups in the last Census. The sample design is stratified multi-stage sampling with two

stages in the self-representing (SR) urban areas and three or four stages in the non-self-representing (NSR) rural areas of the design. The sample selection in the initial stages is with probability proportional to population size and that in the last stage, where dwellings are selected from clusters, being systematic. The selected clusters are assigned six rotation numbers independently within each stratum. In any survey month one-sixth of the households have been in the survey from 1 to 6 months. Thus the entire sample is divided into six equally representative sub-samples of equal sizes [4]. The rotation numbers for six rotation groups can be converted to number of "months in the survey" by a simple transformation.

The adjustment of weight for non-response is done for the entire sample in balancing units by ratio of households in the sample to responding households. In the NSR areas each primary sampling unit (PSU) is divided into two balancing units consisting of urban and rural parts. In the SR areas of the design, strata (called sub-units) form balancing units. The number of balancing units thus exceeds 900 in NSR areas and 800 in SR areas.

In order to evaluate the rotation group bias in the LFS estimates, with and without adjustment, data on non-response rates  $(1-R_i)$  and  $\bar{Y}_{1i}$  and  $\bar{Y}_{2i}$ , proportions for the characteristics "employed" and "unemployed" for respondents and non-respondent respectively in twelve surveys in 1981 are presented and analyzed in Section 3. The "months in the survey" represents number of months (including the current month) a rotation group is in the survey. No data on response biases,  $\bar{\beta}_i$ , are presented.

### 3. ANALYSIS OF LFS DATA

Table 1 shows average non-response rates,  $(1-R_i)$ , by months in the survey for calendar months in 1981. It can be seen that the rates differ substantially between the two areas and between months in the survey for a given area. In both the areas and at Canada level, non-response rates are high in the first month, decrease substantially in the second month and decrease slowly over the succeeding months. The high non-response rates in the first month are contributed by "temporary absent" and "no-one-at home" type households. In the later months the rates reduce due to interviewer's knowledge about the

best time to call on these households. The rates are higher in SR areas, especially apartments (not shown in the table) as compared to NSR areas. During processing, for approximately 1/2% households data are carried forward from the previous month. The non-response rates presented in the tables are obtained by considering those households as respondent. It may be noted that difference of rates from their mean ( $R_i - \bar{R}$ ), is negative in the first and in some cases in the second month in the survey and positive in the following months. The mean rate  $\bar{R}$  is approximately equal to  $R_2$ . Thus from (4) relative bias for first month in the survey is expected to increase, if  $(\bar{Y}_{1i} + \bar{\beta}_i)$  and population mean  $\bar{Y}_i$  are assumed constant; for months 3 to 6, the relative bias is expected to decrease after adjustment of weight for non-response.

Table 2 shows estimated proportions,  $\hat{Y}_{1i}$  and  $\hat{Y}_{2i}$ , of employed and unemployed heads of households by months in the survey for respondent and non-respondent households respectively. The estimates were obtained from LFS longitudinal files for the period March - August 1976 and are based on unweighted counts. The data on non-respondents, who responded at least once during the six month period, were obtained from months in which they responded. Non-respondent households tend to have greater proportion of employed heads and lesser proportion of unemployed heads as compared to respondent households. It is known that the difference of proportions between respondents and non-respondents for employed persons tends to be 0.10 and that for unemployed persons tends to be about 0.005, the signs of differences remaining the same. No particular trend over months in the survey can be observed in the proportions of employed and unemployed heads among respondent and non-respondent households.

The contribution of the first month to the first component is negative in all calendar months for both unemployed and employed. This indicates that the bias for the first month in the survey is expected to increase after adjustment for non-response.

The analysis in sections 2 and 3 isolates rotation groups as groups considered for non-response adjustment. For real data, the same relative changes may not be seen due to impact of differential response rates in other groups and changes in magnitude of  $\bar{Y}_{1i}$  and  $\bar{\beta}_i$  during the six month period. In section 5,



we analyze the impact of non-response adjustment by rotation groups on rotation group bias in the LFS estimates and attempt to explain the results on the basis of the model.

It may be noted that non-response adjustment in the present weighting of LFS data is done within balancing units which are much smaller than NSR and SR areas within a province. Thus the estimates of rotation group bias based on the present weighting and non-response adjustment are corrected for differential non-response rates between the two areas but not for those between rotation groups.

#### 4. WEIGHT-ADJUSTMENT BY ROTATION GROUPS

The LFS final weight is composed of five factors: (1) mathematical weight, (2) rural-urban factor, (3) cluster sub-weight (4) balancing factor and (5) age-sex factor. The mathematical weight for a household is the inverse of overall sampling ratio for the household, based on the sample design. Within each province the weight is the same within urban (SR) and rural (NSR) strata except in a few cases, resulting in twenty areas at Canada level with the same mathematical weight. The cluster sub-weight is the inverse of sampling ratio within a cluster. The balancing factor adjusts the weight for non-response and age-sex factor is a ratio adjustment factor based on projected population within age-sex groups at province level.

As explained in section 2, adjustment of weight for non-response is done within balancing units for the sample of households. For the evaluation of impact of weight adjustment by rotation groups, it was decided to use progressively smaller areas (as balancing units) starting with rotation groups at province level. The adjustment of final weight within rotation groups in these areas was done by multiplying by adjustment factors:

$$R_{H(i)} = \frac{\text{respondent households in the sample}}{\text{respondent households in rotation group (i)}}$$

$$R_{P(i)} = \frac{\text{respondent persons in the sample}}{\text{respondent persons in rotation group (i)}}$$

The first factor weights up the estimate of households within a rotation group in a balancing unit to the level of sample of respondent households. The balancing factor weights it up to the level of sample of households within the balancing unit. The second factor, based on the count of respondent persons weights up the estimates to the level of the entire sample of respondent persons and thus corrects the estimates for different household sizes or coverage of persons within households. It is known that non-respondent households tend to have smaller sizes as compared to respondent households. The difference in non-response rates between rotation groups may result in differences in average household sizes.

If  $\hat{Y}(i)$  is estimates total of  $i_{th}$  rotation group and  $Y(i)$ , true value of  $i_{th}$  group total, then the estimate of relative bias of estimated total of  $i_{th}$  rotation group is given by

$$B_y(i) = \frac{\hat{Y}(i) - Y(i)}{Y(i)} ; i = 1, 2, \dots, 6. \quad (6)$$

Since  $Y(i)$ 's are not known and can be assumed to be approximately equal (since rotation groups have equal expected sizes at large area level)  $\hat{Y}(\cdot)$ , the mean of six rotation group total estimates can be used in place of  $Y(i)$ . The rotation group bias index for  $i_{th}$  rotation group is given by

$$I_y(i) = \frac{\hat{Y}(i)}{\hat{Y}(\cdot)} \cdot 100 = 1 + \beta_y(i) \cdot 100 \quad (7)$$

It may be noted that, since the mean of estimates of six rotation group totals is used instead of true values,  $I_y(i)$  may be biased but is useful as a measure for evaluation of difference in relative biases between rotation groups for various sub-groups of the population and adjustment of weight based on household and person counts. Similarly,  $P_y(i)$ , the rotation group bias of population estimate can be defined for individual rotation groups. The values of the index  $I_y(i)$  above 100 indicate positive relative bias and the values below 100 indicate negative relative bias. Similarly, the index  $I_p(i)$  can be interpreted.



## 5. ANALYSIS OF DATA ON ROTATION GROUP BIAS INDEX

In the following tables data on rotation group bias index for population and labour force status categories by type of area and age-sex groups are presented and analyzed. The index values are obtained by using final weights and the same adjusted for non-response by rotation groups using each of the two factors based on household and person counts. A comparison of index values based on adjusted and unadjusted weights is used in evaluation of impact of weight adjustment on estimates of rotation group bias. The adjustment of weight by rotation groups, using household counts, was done at province level. Thus the final weights for households in the six rotation groups in each province were multiplied by adjustment factors  $R_H(i)$ ;  $i = 1, 2, \dots, 6$ . Similarly, the adjustment based on count of persons was done at province level by factors  $R_P(i)$ ;  $i = 1, 2, \dots, 6$ . In order to evaluate the impact of these adjustments on estimates of population we present Table 3 showing rotation group bias index for population estimates by type of area and months in the survey for twelve surveys in 1981. The index values based on unadjusted weight indicate that there is relative underestimation of persons in the first and the sixth month in both SR and NSR areas. The index values based on weight adjustment using household counts show some improvement in bias; however, this adjustment assumes that household size is the same in six rotation groups. The index values based on weight adjustment using counts of respondents are closer to 100.0 in both the areas, as compared to those based on household adjustment. Thus, the adjustment based on count of persons seems to correct the estimates for differential bias better than the adjustment based on household counts. The higher index values in earlier months and lower in later months could be due to changes in size of non-responding households by month in the survey.

Tables 4 and 5 present data on average index values by type of area and age-sex groups for twelve surveys in 1981. Index values by type of area based on unadjusted weight indicate that relative bias of estimates of unemployed tends to be positive in the first two months and shows a decreasing trend in the later months. Those for employed and in labour force tend to be negative in the first month and positive in the following months. Data on index values by age-sex groups show similar trends as those by type of area.

The adjustment of weight for non-response based on household counts tends to increase the index values in the first month and also fifth and sixth months. The index values in other months tend to decrease. This is true for index values for labour force status by type of area and age-sex groups. The increase in index values in the first month can be attributed to lower than average response rates and the decrease in index values in the following months to higher than average response rates. The decrease in the last two months can not be explained on the basis of higher than average response rates if  $(\bar{Y}_{1i} + \bar{\beta}_i)$  is assumed constant.

The adjustment of weight for non-response based on count of persons tends to increase the index values in the first month and decrease the index values in the third to sixth month. The index values for the first month based on adjustment using count of persons tend to be greater than those based on household adjustment. The adjustment based on count of persons seems to correct the estimates for differential response between rotation groups. The response rates are low in the first month resulting in increase in relative bias after adjustment. The decrease in the relative bias in the third to sixth month seems to be due to lower than average response rates at household level, corrected for differential household size between rotation groups.

## 6. SUMMARY AND CONCLUDING REMARKS

This paper considers a model which decomposes overall bias into three components, showing the contribution due to differences in response rates, response biases and characteristics of respondents and non-respondents between groups of a sample. Rotation groups can be considered as a particular case of these groups in which adjustment of weight for non-response can be done separately. The model also shows contribution of various factors to rotation group bias.

If response rates differ between rotation groups, and the proportion of a characteristic for respondents and the associated response bias is equal for all rotation groups, non-response adjustment by rotation groups does not change the bias of estimates. However, rotation group bias can increase or decrease, according as response rate is lesser or greater than the mean response rate. This is corroborated by data on index values before and after adjustment of weight, based on count of persons.

It is proposed to analyze index values for labour force status and other characteristics for larger data sets and to study the impact of differences in average household sizes between rotation groups and respondent and non-respondent households on estimates of rotation group bias. The contribution of differential response rates and response biases to rotation group bias, after adjustment for non-response by rotation groups, will also be analyzed.

### ACKNOWLEDGEMENTS

The author would like to thank R. Vettore and R. Barnes for the development of computer programs and the referees for helpful comments.

### REFERENCES

- [1] Bailar, B.A., (1975), The Effect of Rotation Group Bias on Estimates from Panel Surveys, JASA, Vol. 70, pp 23-30.
- [2] Bailar, B.A., Bailey, L. and Corby, C., (1978), A Comparison of Some Adjustment and Weighting Procedures for Survey Data, Survey Sampling and Measurement, Academic Press, pp 175-198.
- [3] Platek, R., Singh, M.P. and Tremblay, V., (1978), Adjustment for Non-Response in Surveys, Survey Methodology, Vol. 3, No. 1, pp 1-24.
- [4] Statistics Canada (1976), Methodology of the Canadian Labour Force Survey, Catalogue 71-526, Occasional.
- [5] Tessier, R. and Tremblay, V., (1976), Findings on Rotation Group Biases, Internal Technical Memorandum, Statistics Canada.
- [6] Thomsen, I., (1973), A Note on the efficiency of Sub-class Means to Reduce the Effects of Non-response when Analyzing Survey Data, Statistical Review, Published by the National Central Bureau of Statistics, Stockholm, Sweden, Vol. 11, No. 4.

- [7] Williams, W.H. and Mallows, C.L., (1970), Systematic Biases in Panel Surveys Due to Differential Non-Response, JASA, Vol. 65, No. 331.
- [8] Woltman, H. and Bushery, J., (1975), A Panel Bias Study in the National Crime Survey, presented at Annual ASA Meeting.

E 1. % Non-Response Rates for Households by Months in Survey and Type of Area (1981)

Months	Type of Area		Canada†
	NSR†	SR	
1	6.6	7.9	7.3
2	4.0	4.6	4.4
3	3.5	4.4	3.9
4	3.5	4.1	3.8
5	3.2	3.8	3.6
6	3.1	3.6	3.4
Age No. of Households	26,707	28,645	55,352
cluding special areas			

E 2. Estimated Proportions of Employed and Unemployed Heads in Respondent and Non-Respondent Households

	Respondents $\hat{\bar{Y}}_{1i}$		Non-Respondent $\hat{\bar{Y}}_{2i}$		$\hat{\bar{Y}}_{1i} - \hat{\bar{Y}}_{2i}$	
	Employed	Unemployed	Employed	Unemployed	Employed	Unemployed
	0.6893	0.0383	0.7839	0.0335	-0.0946	0.0048
	0.6962	0.0344	0.7841	0.0321	-0.0879	0.0023
	0.7006	0.0311	0.7851	0.0300	-0.0845	0.0011
	0.7006	0.0364	0.7877	0.0281	-0.0871	0.0083
	0.6972	0.0317	0.7821	0.0317	-0.0849	0.0000
	0.6927	0.0331	0.7767	0.0320	-0.0840	0.0011
Age	0.6961	0.0342	0.7833	0.0311	-0.0872	0.0031



TABLE 3. Rotation Group Bias Index for Population by Type of Area

Weight	Type of Area	Month in the Survey					
		1	2	3	4	5	6
Unadjusted	SR	97.0	101.1	101.2	100.6	100.2	99.8
	NSR	97.7	101.0	100.8	100.9	100.2	99.8
Household adjusted	SR	98.7	98.7	99.4	100.0	101.1	100.0
	NSR	99.3	98.6	99.0	100.3	101.1	100.0
Population adjusted	SR	100.4	100.5	100.2	99.7	99.6	99.8
	NSR	100.9	100.3	99.8	99.9	99.6	99.8

TABLE 4. Rotation Group Bias Index by Type of Area (1981)

Weight	Character-istics	Type of Area	Month in the Survey					
			1	2	3	4	5	6
Unadjusted	Employed	SR	99.9	101.0	101.3	100.7	100.4	99.8
		NSR	96.8	100.9	100.6	101.2	100.7	99.8
	Unemployed	SR	99.1	102.6	101.3	100.4	97.7	98.8
		NSR	103.3	101.5	101.4	99.8	96.5	97.8
	In LF	SR	97.0	101.1	101.3	100.7	100.2	99.8
		NSR	97.3	100.9	100.7	101.1	100.3	99.8
Household adjusted	Employed	SR	98.6	98.5	99.5	100.1	101.2	102.0
		NSR	98.3	98.4	98.7	100.6	101.6	102.0
	Unemployed	SR	100.8	100.3	99.5	99.8	98.5	101.0
		NSR	104.9	99.2	99.6	99.2	97.3	99.8
	In LF	SR	98.7	98.7	99.5	100.1	101.0	102.0
		NSR	98.9	98.4	98.8	100.5	101.2	102.0
Population adjusted	Employed	SR	100.2	100.3	100.4	99.7	99.7	99.8
		NSR	100.0	100.2	99.6	100.2	100.1	99.8
	Unemployed	SR	102.4	102.1	100.4	99.4	97.1	98.8
		NSR	106.4	100.8	100.5	98.9	96.0	97.8
	In LF	SR	100.4	100.5	100.4	99.7	99.5	99.8
		NSR	100.6	100.2	99.7	100.1	99.8	99.8

TABLE 5. Rotation Group Bias Index by Age-Sex Groups (1981)

Weight	Character- istics	Age-Sex Group	Month in the Survey					
			1	2	3	4	5	6
Right adjusted	Employed	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
		F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
		M 25+	97.0	101.4	101.3	100.7	100.1	99.5
		F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	Unemployed	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
		F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
		M 25+	98.0	102.1	101.6	100.3	98.0	100.1
		F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	In LF	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
		F 15-24	96.8	100.0	100.7	100.7	101.0	100.8
		M 25+	97.1	101.4	101.3	100.7	100.0	99.5
		F 25+	97.1	101.3	101.4	101.0	100.1	99.1
Household adjusted	Employed	M 15-24	98.2	97.2	98.8	100.3	101.9	103.5
		F 15-24	97.6	97.2	99.3	100.3	102.1	103.5
		M 25+	98.7	98.9	99.4	100.1	101.0	101.8
		F 25+	98.6	98.8	99.3	100.3	101.3	101.6
	Unemployed	M 15-24	102.6	100.0	99.3	100.1	97.7	100.3
		F 15-24	104.2	100.3	96.0	98.3	100.8	100.4
		M 25+	99.6	99.8	99.8	99.7	98.8	102.3
		F 25+	101.8	100.0	102.6	99.8	96.1	99.6
	In LF	M 15-24	98.9	97.6	98.9	100.3	101.3	103.0
		F 15-24	98.5	97.6	98.8	100.0	101.9	103.2
		M 25+	98.8	99.0	99.5	100.1	100.8	101.0
		F 25+	98.8	98.8	99.6	100.3	101.0	101.5
Population adjusted	Employed	M 15-24	99.9	99.0	99.7	100.0	100.5	101.0
		F 15-24	99.3	99.1	100.2	100.0	100.6	101.0
		M 25+	100.4	100.7	100.3	99.7	99.5	99.3
		F 25+	100.3	100.6	100.2	100.0	99.9	99.1
	Unemployed	M 15-24	104.2	101.7	100.1	99.7	96.3	98.0
		F 15-24	105.8	102.0	96.8	98.0	99.4	98.0
		M 25+	101.2	101.6	100.7	99.4	97.4	99.8
		F 25+	103.4	101.7	103.5	99.5	94.8	97.1
	In LF	M 15-24	100.5	99.4	99.8	99.9	99.8	100.5
		F 15-24	100.2	99.4	99.7	99.7	100.4	100.6
		M 25+	100.4	100.8	100.3	99.7	99.4	99.3
		F 25+	100.5	100.6	100.4	99.9	99.5	98.9

COMPUTERIZATION OF COMPLEX SURVEY ESTIMATES<sup>1</sup>M.A. Hidioglou<sup>2</sup>

Survey data collected by statistical agencies is most likely to be processed through to the tabulation stage by these agencies. The computer programs associated with this processing are also most likely tailored to the particular design and variables used. The statistics computed from such surveys typically range from simple descriptive totals and means to those required for analytic studies such as comparison of domains, regression analysis and contingency tables analysis. This paper describes a computer program which computes these statistics and their associated sampling errors for commonly used sampling designs.

## 1. INTRODUCTION

A variety of statistics are computed for survey data which often arise from large, complex national and regional surveys. The statistics computed from such surveys typically range from simple descriptive totals and means to those required for analytic studies such as comparison of domains, regression analysis, and contingency tables analysis. Domain estimation refers to the estimation of statistics for subgroups of the population of interest which are not explicitly provided for in the design. Yates (1960) contains considerable material on the estimation of domain means and their differences. Hartley (1959) and Rao (1975) provide an excellent account of the methodology used for domain estimation. The variance estimators associated with the domain estimators are easy extensions of variance estimators for simple statistics. This is not, however, the case for more complex statistics. The estimation of regression equations from survey data presents several problems; for example, the definition of the regression equations, the identification of the population for which inferences are desired, and the variance estimation for the regression coefficients (see Konijn (1962), Kish and Frankel (1974) and

---

<sup>1</sup> Presented at the Annual Meetings of the American Statistical Association, Detroit, August 1981.

<sup>2</sup> M.A. Hidioglou, Business Survey Methods Division, Statistics Canada.

Fuller (1975). The testing of hypotheses for contingency tables given survey design considerations have been studied by Nathan (1969, 1972), Rao and Scott (1981), Garza-Hernandez and McCarthy (1962) and Koch, Freeman and Freeman (1975) to name a few.

Survey data collected by statistical agencies is most likely processed through to the tabulation stage by these agencies. The computer programs associated with this processing are also most likely tailored to the particular design used. It is quite possible that computer programs used to produce estimates of totals (say) and their associated variances must be developed from scratch every time that a new survey design is introduced. This is time consuming, expensive, tedious and in some sense repetitive. Use of statistical software packages such as SPSS or SAS may be considered as an alternative. These packages may be readily used to produce weighted estimates. However, the variances that they compute do not take sample design factors such as stratification and clustering into account unless they are programmed to do so. A user must therefore be fairly familiar with the language used by these packages if he wants to obtain proper variance estimates for survey estimates.

Recently, there have been attempts to develop programs which compute variances for a general class of designs. Some of these programs are STDERR by Shah (1974), SURREGR by Holt (1975), SUPER CARP and MINI CARP by Hidiroglou, Fuller and Hickman (1980). These programs basically require the specification of the estimator to be used and the variables to be analysed. It will be assumed that the data sets that these programs are being applied to have been edited and that missing observations have been imputed. In this paper, SUPER CARP and MINI CARP will be described. SUPER CARP can be used to construct estimated totals, ratio estimates, the difference of ratio estimates and contingency tables tests for multistage stratified samples. It contains a number of regression procedures appropriate for data observed subject to response (Measurement) error. Covariance matrices can be estimated for sub-population means, and totals and for stratum means and totals. MINI CARP is a smaller program which differs from SUPER CARP in that it does not contain and of SUPER CARP's regression procedures. A comparison of the capabilities of the two programs is given in Table 1.



TABLE 1. Capabilities of SUPER CARP (S) and MINI CARP (M)

Multivariate Estimate of	For		
	Entire Population	Individual Strata	Sub- population
<u>Simple Parameters</u>			
. Means	S,M	S,M	S,M
. Totals	S,M	S,M	S,M
. Ratios	S,M	S,M	S,M
. Difference of Ratios	S,M	S,M	S,M
. Proportions	S,M	S,M	S,M
<u>Complex Parameters</u>			
		<u>Tests</u>	
. Weighted Least Squares	S	. Regression	
. Weighted Errors-in-the Variables (Known & Estimated error covariances)	S	Coefficient	S
		. Goodness-of-fit	S,M
		. Independence for Two-Way Table	S,M

## 2. GENERAL DESCRIPTION

### 2.1 Notation

In general, SUPER CARP and MINI CARP can accept data from a multistage stratified design. Assuming that the design has  $s$  stages, a  $g$  dimensional data vector is read in for each observation. We denote this data vector as

$$(z_{hi_s1}, z_{hi_s2}, \dots, z_{hi_sg}),$$

where  $h = 1, 2, \dots, L$  denotes strata;  $i = (i_1, i_2, \dots, i_s)$  represent the stages;  $i_1 = 1, 2, \dots, n_h$  represents the first stage identification;  $i_2 = 1, 2, \dots, n_{hi_1}$  represent the second stage identification; ... ;  $i_s = 1, 2, \dots, n_{hi_{s-1}}$  represents the last  $s$ -th identification.  $z_{hi_s k}$  is the  $hi_s$ -th observation for the  $k$ -th variable of interest. Weights associated with the  $hi_s$ -th observation will be referred to as  $w_{hi_s}$ . These weights would be inversely proportional to the selection probabilities of each ultimate sampled



unit. The specification of the variables to be used in the analysis (be it total or ratio estimation or regression estimation) is done by using a selection vector  $y = (v_1, v_2, \dots, v_{p+1})$  where  $1 \leq v_k \leq g$  for  $k = 1, 2, \dots, p + 1$ . Given that the type of analysis and the identification of the variables has been decided upon, let the chosen vector for the  $h_{i_s}$ -th observation be

$$(Y_{h_{i_s}}, X_{h_{i_s}1}, X_{h_{i_s}2}, \dots, X_{h_{i_s}g}),$$

where  $Y$  denotes the dependent variable and  $X$  denotes the independent variables if regression analysis is specified. Note that  $v_1$  is always the index for the dependent variable in the case of regression. For other types of analyses, the ordering within the selection vector is not important.

## 2.2 Types of Computations

The simple statistics and a partial list of the regression options available in the program are outlined. A complete description of all the available options is written up in the SUPER CARP or MINI CARP manuals (1980).

(i) Total Estimator, e.g.

$$\hat{X}_{(k)} = \sum_h \sum_{i_1} \dots \sum_{i_s} w_{h_{i_s}} X_{h_{i_s}(k)}, \quad k = 1, 1, 2, \dots, p.$$

The estimated covariance matrix for

$$\hat{X} = \{\hat{X}_{(1)}, \hat{X}_{(2)}, \dots, \hat{X}_{(p)}\} \text{ is}$$

$$v_1(\hat{X}) = \sum_{h=1}^L (n_h - 1)^{-1} n_h (1 - f_h) \sum_{i_1=1}^{n_h} (\hat{d}_{h_{i_1} \cdot} - \hat{d}_{h..})^T (\hat{d}_{h_{i_1} \cdot} - \hat{d}_{h..}) \quad (2.2.1)$$

where

$$\hat{d}_{hi_1.} = \{\hat{d}_{hi_1(1)}, \hat{d}_{hi_1(2)}, \dots, \hat{d}_{hi_1(p)}\}$$

$$\hat{d}_{hi_1(k)} = \sum_{i_2=1}^{n_{hi_1}} \dots \sum_{i_s=1}^{n_{hi_{s-1}}} w_{hi_s} x_{hi_s(k)}$$

$$\hat{d}_{h..} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{hi_1.}.$$

Note that the above variance formula may be applied to pps schemes with and without replacement. For with replacement schemes, only the first stage variance needs to be computed (Des Baj, 1968, pg. 120) and the correction factors  $f_h$  are set to zero. In large scale surveys, it is often assumed that the first stage clusters have been selected without replacement even though the actual selection scheme may have been without replacement. This assumption inconjunction with small sampling fractions implies that resulting variance is fairly close to the one which would have been obtained by taking all stages and selection procedure into account. If the sampling fractions are not negligible at each stage and that the sampling has been performed using without replacement S.R.S. at each stage, Des Raj's rule (1966) can be used to advantage to compute each stage component of covariance. The covariance matrix accounting for  $s$  stages is:

$$v(\hat{X}_{\sim}) = \sum_{r=1}^s v_r(\hat{X}_{\sim})$$

where for  $r \geq 2$

$$v_r(\hat{X}) = \sum_{h=1}^L \sum_{i_1=1}^{n_h} \dots \sum_{i_{r-1}=1}^{n_{hi_{r-2}}} \left[ \begin{matrix} r-2 \\ \pi \\ j=0 \end{matrix} \frac{n_{hi_j}}{N_{hi_j}} \right]$$

(2.2.2)

$$\times n_{hi_{r-1}} (n_{hi_{r-1}} - 1)^{-1} (1 - f_{hi_{r-1}})$$

$$\times \sum_{i_r} (\hat{d}_{hi_r}(\cdot) - \hat{\tilde{d}}_{hi_{r-1},\cdot}(\cdot))^T$$

$$\times \hat{d}_{hi_r}(\cdot) - \hat{\tilde{d}}_{hi_{r-1},\cdot}(\cdot)$$

where

$$f_{hi_{r-1}} = n_{hi_{r-1}} N_{hi_{r-1}}^{-1}, n_{hi_0} = n_h, N_{hi_0} = N_h,$$

$$\hat{d}_{hi_r}(\cdot) = \hat{d}_{hi_r}(1), \hat{d}_{hi_r}(2), \dots, \hat{d}_{hi_r}(\rho)$$

$$\hat{d}_{hi_r}(k) = \sum_{i_{r+1}} \dots \sum_{i_s} w_{hi_s} x_{hi_s}(k)$$

$$\hat{\tilde{d}}_{hi_{r-1},\cdot}(\cdot) = n_{hi_r}^{-1} \sum_{i_r} \hat{d}_{hi_r}(\cdot)$$

The variance estimation for an  $r$ -stage design can therefore be done by estimating the components at each stage ( $v_r(\hat{X})$ ) and summing them up. This can be done by passing over the data set  $r$  separate times. The first time around, strata and first stage units are read into the program to give  $v_1(\hat{X})$ . The second time around, the original primary sampling units are read into the program as "strata" and the secondary units are identified as clusters to give  $v_2(\hat{X})$ . The  $r$ -th time around, the original ( $r \geq 2$ ) ( $r-1$ )-th stage units are read into the program as "strata" and the  $r$ -th stage units are identified as clusters to give  $v_r(\hat{X})$ .

On each pass a sampling rate  $g_{h_{i_{r-1}}}$  must be read in for the  $h_{i_{r-1}}$ -th unit where

$$g_{h_{i_{r-1}}} = 1 - \left[ \frac{r-2}{\pi} \frac{n_{h_{i_{r-1}}j}}{N_{h_{i_{r-1}}j}} \right] \left[ 1 - \frac{n_{h_{i_{r-1}}-1}}{N_{h_{i_{r-1}}-1}} \right] .$$

Using this procedure, the program will be computing  $v_r(\hat{X})$  in the format given by  $v_1(\hat{X})$ .

If the sampling factors are not negligible at each stage and that sampling has been performed using without replacement p.p.s. schemes at each stage, the variance expression at each stage must take into account joint selection probabilities. SUPER CARP and MINI CARP do not compute joint selection probabilities. For the case where two units per stratum have been selected without replacement and unequal probability, the variance of the estimator for total can be obtained using formula (2.2.1) with a correction factor for each stratum which includes the joint probabilities of selection. This correction factor is given by

$$f_h = \frac{2\pi_{h12} - \pi_{h1}\pi_{h2}}{\pi_{h12}} , h=1, 2, \dots, L$$

where  $\pi_{h12}$  is the joint probability of selection for the selected units 1 and 2. If  $n_h \geq 2$  and that the joint probabilities of selection are not available, an approximation to the without replacement variance has been given by Gray

(1975). Gray shows that the variances of an unequal without replacement sample may be partitioned into a "with replacement" variance component times a finite population correction factor which depends on the joint probabilities. This correction factor has been found to be roughly equal to one minus the inverse of the sampling fraction for populations which have more than 15 elements within each stage. Using Gray's approximation, variances for multistage unequal without replacement schemes can be computed.

If domain estimation is required for some of the variables, a new variable  $d^{Y_{hi_s}}(k)$  is defined for all elements in the population, where

$$d^{Y_{hi_s}}(k) = \begin{cases} Y_{hi_s}(k) & \text{if the } hi_s\text{-th element belongs} \\ & \text{to the domain } d \text{ (say } D_d) \\ 0 & \text{otherwise} \end{cases}$$

An alternative way of defining  $d^{Y_{hi_s}}(k)$ , is

$$d^{Y_{hi_s}}(k) = d^{a_{hi_s}} Y_{hi_s}(k) \text{ where}$$

$$d^{a_{hi_s}} = \begin{cases} 1 & \text{if the } hi_s\text{-th element belongs to } D_d \\ 0 & \text{otherwise} \end{cases}$$

Note that if  $\hat{Y}$  and  $v(\hat{Y})$  are unbiased for  $Y$  and  $v(Y)$  respectively, then the corresponding domain estimators  $\hat{d}^{\hat{Y}}$  and  $v(\hat{d}^{\hat{Y}})$  are unbiased for  $d^Y$  and  $V(d^Y)$ . The standard formulae for  $\hat{Y}$  and  $v(\hat{Y})$  can now be applied to the "synthetic" variables  $d^{Y_{hi_s}}$ . Stratum totals can be computed individually by treating the strata as classification variables.

## (ii) Ratio Estimator

The vector  $\{Y_{hi_s}(1), X_{hi_s}(1), \dots, Y_{hi_s}(p), X_{hi_s}(p)\}$  is used in the analysis and the estimated ratios are:

$$\hat{R}(t) = \hat{X}_{(t)}^{-1} \hat{Y}_{(t)}, t = 1, 2, \dots, p ;$$



where  $\hat{Y}_{(t)}$  and  $\hat{X}_{(t)}$  are of the form given in the previous section. The estimated covariance matrix for  $\hat{R} = \{\hat{R}(1), \hat{R}(2), \dots, \hat{R}(p)\}$  is as given in the previous section with

$$\hat{d}_{hi_r}(t) = \hat{X}_{(t)}^{-1} \sum_{i_{r+1}} \dots \sum_{i_s} w_{hi_s} \{Y_{hi_s}(1) - \hat{R}(t) X_{hi_s}(t)\}; t = 1, \dots, p.$$

The ratio estimator can be used for computing the mean for each variable of interest by setting all X-variables to 1. Domain means can be computed by using  $d_{hi_s}^Y(t)$  in the place of  $Y_{hi_s}(t)$  and  $d_{hi_s}^X$  in the place of  $X_{hi_s}(t)$ . If subpopulation proportions of Y for a domain  $D_d$  are required, the numerator of the ratio is the sum of weighted  $d_{hi_s}^Y(t)$  and the denominator is the sum of weighted  $Y_{hi_s}(t)$ . The estimated ratio for two variables defined over a domain  $D_d$  may similarly be obtained. Stratum proportions and ratios may be computed with the strata serving as the classification variables.

### (iii) Regression Estimation

Some considerable attention has been paid recently to regression concepts in survey sampling. There are several explanations for this. First, there is an increased emphasis on analytic surveys, with partly unresolved questions of proper weighting of observations. Secondly, modeling in general, especially in the regression context, has attracted widespread interest, as well as criticism, as a tool in making survey estimates. SUPER CARP properly weights the observations and computes the variances of the estimated regression coefficients using a method given by Fuller (1975).

The regression coefficients estimated from a stratified cluster sample are given by

$$b_{\sim} = (X'_{\sim} W X_{\sim})^{-1} X'_{\sim} W Y_{\sim}$$

where the (rs)-th element of  $(X_n' W X_n)$  is

$$\sum_{h=1}^L \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} X_{hi_1i_2s} W_{hi_1i_2}$$

and the r-th element of  $X_n' W Y_n$  is

$$\sum_{h=1}^L \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} X_{hi_1i_2} W_{hi_1i_2}.$$

The estimated covariance matrix of  $b_n$  is computed as

$$v(b) = (X_n' W X_n)^{-1} \hat{G}_n (X_n' W X_n)^{-1},$$

where the (rs)-th element of  $\hat{G}_n$  is

$$\hat{g}_n(r,s) = \frac{n-1}{n-p} \sum_{h=1}^L \frac{n_h(1-f_h)}{(n_{h-1})} \sum_{i_1=1}^{n_h} (\hat{d}_{hi_1.r} - \bar{d}_{h..r}) \times (\hat{d}_{hi_1.s} - \bar{d}_{h..s})$$

where

$$\hat{d}_{hi_1i_2r} = X_{hi_1i_2r} \hat{v}_{hi_1i_2} W_{hi_1i_2},$$

$$\hat{v}_{hi_1i_2} = Y_{hi_1i_2} - \sum_{r=1}^p \hat{b}(r) X_{hi_1i_2r},$$

$$\hat{d}_{hi_1.r} = \sum_{i_2=1}^{n_{hi_1}} \hat{d}_{hi_1i_2r},$$

$$\bar{d}_{h..r} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{hi_1.r},$$

$$n = \sum_{h=1}^L \sum_{i_1=1}^{n_h} n_{hi_1}.$$

The variance estimation procedure is based on an asymptotic Taylor expansion of the sample regression coefficient vector. This method has several advantages over the Balanced Repeated Replication and Jack-Knife Replication methods. Firstly, it is relatively easy to program, and it can be adopted to multistage sample designs. Secondly, no restrictions are placed on the sample design (two replicates per stratum, for instance) and the assumptions used require some well-behaved moments in the population of interest. Thirdly, it requires the least number of computations.

Data is quite frequently measured with error. Theory for regression models which takes measurement error into account has been given by Fuller (1980a), Fuller (1980b) and Fuller and Hidiroglou (1978). SUPER CARP also has the flexibility to compute tests of hypothesis for any subsets of the regression parameters.

#### (iv) Contingency Tables

SUPER CARP and MINI CARP perform the goodness-of-fit test and the independence test for data resulting from complex surveys. These two tests take the stratification and the clustering of the design into account. As pointed out by Rao and Scott (1981), practitioners using traditional Pearson chi-square statistics for those two tests, given that there may be serious design effects can be seriously misled.

For the goodness-of-fit test, SUPER CARP and MINI CARP use the modified Wald Statistic given by

$$F_{WG} = [(k-1)d]^{-1} (d-k+2) (\hat{p} - p_0)^T \hat{V}^{-1} (\hat{p} - p_0)$$

where

$\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_{k-1})^T$  is the vector of estimated proportions given in the stratum and cluster configurations,

$p_0 = (p_{01}, p_{02}, \dots, p_{0,k-1})^T$  is the vector of hypothesized proportions,

$\hat{V}$  = the covariance matrix of  $\hat{p}$  given the stratum and cluster configuration,

$k$  = number of categories considered,

$$d = \sum_{h=1}^L (n_h - 1),$$

$L$  is the number of strata in the sample and  $n_i$  is the number of clusters in the  $i$ -th stratum. The covariance matrix  $\hat{V}$  is computed using the methods given for ratio estimation. In large samples,  $F$  is approximately distributed as a central  $F$  with  $k-1$  and  $d-k+2$  degrees of freedom when the null hypothesis is true.

For the test of independence, Fuller (SUPER CARP p. 65-69) has developed a test which takes the design into account. Given that the contingency table which splits the population according to two criteria is made up of  $R$  rows and  $C$  columns, the null hypothesis to be tested is  $H_0: p_{ij} = p_{i+} p_{+j}$  or

$p_{+j} = p_{i+}^{-1} p_{ij}$  . where  $p_{ij}$  =  $ij$ -th cell proportion in the population,

$p_{+j} = \sum_i p_{ij}$  and  $p_{i+} = \sum_j p_{ij}$  .

Given that  $P_{ij|i}$  is defined as  $P_{i+}^{-1} P_{ij}$  and that the corresponding sample estimators are  $\hat{P}_{ij|i} = \hat{P}_{i+}^{-1} \hat{P}_{ij}$ , estimates for  $(P_{+1}, P_{+2}, \dots, P_{+,C-1})$  can be obtained by regressing  $\hat{P}_{ij|i}$  ( $i = 1, 2, \dots, R; j = 1, 2, \dots, C-1$ ) on  $(C-1)$ -dimensional row vectors whose elements are one for the  $j$ -th entry corresponding to  $\hat{P}_{ij|i}$  and zero otherwise. The regression is of a generalized least-squares nature because the  $\hat{P}_{ij|i}$  do not have the same error structure. An estimator for the covariance matrix of the  $P_{ij|k}$ 's, incorporating the sample design, is obtained using the ratio estimator formulae. The test statistic for  $H_0$  is then based on the residual sums of squares for this regression.

### 3. INPUT

In a typical survey situation, the data associated with a given selected unit is characterized by stratum, first stage, second stage up to  $s$ -th stage identification and a sampling weight. The data must be ordered hierarchically with respect to this identification in order to produce estimates of variance which reflect the stratified and clustered of the data.

SUPER CARP and MINI CARP are run using command language specified in numeric codes in fixed card positions. For both programs, there are six mandatory control cards to be input at all times. A number of optional control cards may also be input if more information is required by options specified in the mandatory cards. The mandatory cards are the parameter card, the variable name card, the format card, the screening card, the analysis card and the variable identification card. The parameter card provides overall preliminary information to the program such as, problem identification, number of observations to be read in, input service identification (tape, disk or cards), data identification structure, data output and stratum collapsing controls. The format card specifies the input format for the data as well as its identification and the associated weight. the variable name card assigns chosen names to input data fields in the order that they are read in. The screening card specifies tolerance limits for given variables provided that screening is required. The analysis card specifies the type of analysis to be performed (see table 1). Finally, the variable identification card identifies the variables to be used in the chosen analyses. The optional cards include such



cards as the sampling rate card (sampling rates by stratum can be read in), the errors-in-the variable cards for supplying the program with covariance matrices for variables measured with error, the hypothesis testing card for specifying coefficients in a regression analysis to be tested equal to zero.

#### 4. COMPUTATIONS

##### 4.1 For Means and Corrected Sums of Squares and Cross-Products

The means, corrected sums of squares and cross-products are statistics routinely computed in a survey package. The choice of algorithms for computing these statistics should take into consideration precision, speed and storage requirements. Beaton, Rubin and Barone (1976) have noted that a "concern about highly accurate computation methods must be tempered with a concern for whether the data are accurate enough to make the results meaningful". Different variations of one-pass and two-pass algorithms have been studied by Ling (1974). Ling's conclusion is that there is no universally best algorithm. The best algorithm for a given data set depends on the numbers in that data set. One of his recommendations is to use double precision arithmetic to be beyond the accuracy attainable in single-precision arithmetic. One-pass recursive algorithms should be chosen over the usual one-pass 'desk-machine' method because they have a higher tendency to produce less computational errors. This is especially the case for subroutines programmed in single precision. In SUPER CARP and MINI CARP one-pass recursive algorithms programmed in double precision have been chosen.

##### 4.2 Inversion of Matrices

Matrix inversion is required for regression and contingency table analysis. the choice for inversion algorithms is quite important in packages. This has been reported by Longley's (1967) paper in which he examined the accuracy of some inversion algorithms and found serious computational inaccuracies. He reported that the most accurate results were obtained by using the orthonormalization procedure. Kopitze, Boardman and Graybill (1975) recommend the use of the Cholesky decomposition as an inversting algorithm. They point out that as compared to the Gaussian elimination schemes, it does not require

pivoting to stabilize symmetric positive definite matrices. This means less time for inverting. The Cholesky decomposition does not need much core storage and is easier to program than the Gaussian elimination scheme. One of its other advantages, as Wilkinson's (1965) analysis shows, is that it is quite accurate. Another of its advantages is that it can be used to find eigenvalues for systems of equations of the form  $A \underline{x} = \lambda B \underline{x}$  where  $A$  is a positive matrix and  $B$  is a positive semi-definite matrix. Computations of eigenvalues are required in SUPER CARP for some of the errors-in-the variables regression analyses. It is for this reason and the precision considerations that the Cholesky decomposition has been adopted for inversion purposes in SUPER CARP.

#### 4.3 Stratum Collapsing

If a sampled population is highly heterogeneous and several criteria are available for stratification, it is quite possible that some strata may contain only one cluster. For such strata, it is not possible to estimate the variability. In such cases, the user may request that the one cluster strata be collapsed with neighbouring strata. If such a request is not made, SUPER CARP or MINI CARP exclude with one cluster from variance computations but include them for estimation purposes. The program lists those strata with only one unit. This information may lead the user to collapse those strata in a subsequent pass. If collapsing is to be done, the strata which are to be collapsed should be similar to neighbouring strata. A suggested method for collapsing which is easily amenable to programming is as follows. If a stratum is encountered that contains only one cluster, that stratum is combined with the following stratum in the file sequence. If the last stratum contains only one element, the last stratum is combined with the next to last stratum. A stratum with a sampling rate of one is not collapsed because such a stratum makes no contribution to the between primary component of the sampling variance. Strata with a sampling rate of one should never appear after a stratum with only one cluster. One way to ensure this condition is to place all observations with a sampling rate of one at the beginning of the file sequence. If two strata are collapsed, the resulting sampling rate for the new stratum is computed as a function of the old sampling rates and number of elements in the previous strata.

#### 4.4 Clusters of Size One

If clusters of size one within a stratum at the first stage, collapsing of adjoining strata ensures that variance estimates will be computed. For a multi-stage design, some of the stages may contain single element clusters. For those clusters, no within cluster variation can be computed. There are several ways for handling this situation. One is to assume a zero-variance contribution from those single-element clusters. Another is to collapse them with neighbouring clusters. An alternative is to assume that they contribute a variance equal to the overall within variation of the clusters for which the within variation can be computed. The variance contribution for those stages where some of the clusters are of size one would incorporate this approximation.

#### 5. SOME DESIRABLE FEATURES OF A VARIANCE ESTIMATION PROGRAM

Francis, Heiberger and Velleman (1975) listed criteria useful in evaluating programs in general. In this section, some of the desirable features of a computer program for estimating variance from complex surveys will be listed. These include user's documentation, input controls, printed output and statistical effectiveness. These desirable features will be related to those provided by SUPER CARP and MINI CARP.

User's documentation should consist of a manual which basically tells the user how to use the program. SUPER CARP and MINI CARP both have manuals which explain to the user how to use them. These manuals are structured as follows. They contain an introduction which summarizes the various available statistical options. Data input and command statements used to specify procedures, variables and options are explained and examples are provided to illustrate their use. Since data input and command statements are to be entered in a specific sequence, a flow diagram is provided. The program procedures are described in terms of the formulae used, the numerical techniques employed and some references to the literature.

As stated earlier, the command language used for SUPER CARP and MINI CARP is in the form of code number or alphanumeric codes in fixed card columns. As pointed out by Francis, Heiberger and Velleman (1975), the most computationally efficient command languages employ code number in fixed card columns. The disadvantage of this method is that users may make excessive references to the manual to identify the commands. Procedures and options could have been specified with the addition of a control statement translator which the addition of a control statement translator which would have allowed English like commands. The advantage of this input method is that it is relatively easy to learn. The disadvantage is that the time and effort required for programming this translator can be prohibitive.

The printed output in SUPER CARP and MINI CARP identifies the statistical procedure used and labels the variables used in the analysis. Part of the output refers to the program's version number, name and date it was last updated. This identification can be used to trace and fix bugs in the stated program version. Some informative diagnostic messages are also printed out. These include messages referring to input controls such as attempting to read in more variables than the program has been dimensioned to handle, trying to read too many cluster, improper input format. If some strata contain one cluster, the program will print out list of such strata. If the user requests collapsing of single cluster strata, the resulting strata will be printed out.

SUPER CARP and MINI CARP are written in FORTRAN and in double precision. They can be run on installations that have a FORTRAN compiler with minor modifications to the job control language. They can both be extended to accommodate new statistical procedures. These can be placed in the program in the form of new subroutines which can be connected to existing software in the program.



## REFERENCES

- [1] Beaton, A.E., Rubin, D.B., and Barone, J.L. (1976), The Acceptability of Regression Solutions: Another Look at Computational Accuracy, Journal of the American Statistical Association, 71, 158-168.
- [2] Raj, Des (1968), Sampling Theory. McGraw-Hill Inc.
- [3] Raj, D. (1966), Some Remarks on a Simple Procedure of Sampling Without Replacement, Journal of the American Statistical Association, 61, 393-397.
- [4] Francis, I., Heiberger, R.M., and Velleman, P.F. (1975), Criteria and Considerations in the Evaluation of Statistical Program Packages, The American Statistician, 29, 52-56.
- [5] Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, 37, 117-132.
- [6] Fuller, W.A. and Hidiroglou, M.A. (1968), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.
- [7] Fuller, W.A. (1980a), Properties of Some Estimators for the Errors-in-Variables Model, Annals of Statistics, 8, 407-422.
- [8] Fuller, W.A. (1980b), Estimation of Measurement Error Models from Cluster Samples. Paper presented at the meetings of the American Educational Research Association, Boston, Massachusetts.
- [9] Garza-Hernandez, T. and McCarthy, P.J. (1962), A Test of Homogeneity for a Stratified Sample, Proceedings of the Social Statistics Section of the American Statistical Association, 200-202.
- [10] Gray, C.B. (1975), Components of Variance Model in Multistage Stratified Samples, Survey Methodology, 1, 27-43.



- [11] Hartley, H.O. (1959), Analytic Studies of Survey Data, Instituto de Statistica, Rome, Volume in onore di Corrado Gini.
- [12] Hidioglou, M.A., Fuller, W.A. and Hickman, R.D., (1980), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.
- [13] Hidioglou, M.A., Fuller, W.A., and Hickman, R.D. (1980), MINI CARP, Survey Section, Iowa State University, Ames, Iowa.
- [14] Holt, Mary M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, unpublished report, Research Triangle Institute, Research Triangle Park, North Carolina.
- [15] Kish, L., and Frankel, M.R. (1974), Inference from Complex Samples, Journal of the Royal Statistical Society B, 36, 1-37.
- [16] Konijn, H.S. (1962), Regression Analysis in Sample Surveys, Journal of the American Statistical Association 57, 590-606.
- [17] Kopitze, R., Boardman, T.J. and Graybill, F.A. (1975), Least Square Programs - A Look at the Square Root Procedure, The American Statistician, 29, 64-66.
- [18] Ling, R.G. (1974), Comparison of Several Algorithms for Computing Sample Means and Variances, Journal of the American Statistical Association, 69, 859-866.
- [19] Longley, James (1967), An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User, Journal of the American Statistical Association, 62, 819-841.
- [20] Nathan, G. (1972), On the Asymptotic Power for Tests for Independence in Contingency Tables from Stratified Samples, Journal of the American Statistical Association, 67, 917-920.

- [21] Rao, J.N.K. (1975), Analytic Studies of Sample Survey Data, Survey Methodology Vol 1, supplement, Statistics Canada.
- [22] Rao. J.N.K. (1975), Unbiased Variance Estimation for Multistage Designs, Sankhya C, 37, 133-139.
- [23] Rao, J.N.K. and Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness-of-Fit and Independence in Two-Way Tables, Journal of the American Statistical Association, 76, 221-230.
- [24] Shah, B.V. (1974), STDERR: Standard Errors Program for Sample Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.
- [25] Wilkinson, J.H. (1975), The Algebraic Eigenvalue Problem, Oxford: Clarendon Press, 229-233.
- [26] Yates, F. (1960), Sample Methods for Censuses and Surveys, Charles Griffin and Sons, London, Third Edition.



# SURVEY METHODOLOGY

December 1981

Vol. 7

No. 2

A Journal produced by Methodology Staff, Statistics Canada

## C O N T E N T S

Notes on Inference Based on Data From Complex Sample Designs GAD NATHAN .....	109
The Non-Response Problem J.G. BETHLEHEM and H.M.P. KERSTEN .....	130
On the Variances of Asymptotically Normal Estimators From Complex Surveys DAVID A. BINDER .....	157
An Overview of Canadian Health Statistics: Past, Present and Future LORNE ROWEBOTTOM .....	171
Models for Estimation of Sampling Errors P.D. GHANGURDE .....	177

















Préparé par les méthodologistes de Statistique Canada.

TABLE DES MATIÈRES

109	L'inférence statistique basée sur des plans d'échantillonnage complexes GAD MATHAN .....
131	Le problème de la non-réponse J.G. BETHLEHEM et H.M.P. KERSTEN .....
162	Les variances d'estimateurs asymptotiquement normaux basés sur des enquêtes complexes DAVID A. BINDER .....
179	La statistique de la santé au Canada: rétrospective et jalons pour l'avenir LORNE ROWEBOTTOM .....
187	Modèles d'estimation des erreurs d'échantillonnage P.D. GHANGURDE .....



- [23] Rao, J.N.K. et Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness-of-Fit and Independence in Two-Way Tables. Journal of the American Statistical Association, 76, 221-230.
- [24] Shah, B.V. (1974), STDERR: Standard Errors Program for Sample Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.
- [25] Wilkinson, J.H. (1975), The Algebraic Eigenvalue Problem. Oxford: Clarendon Press, 229-233.
- [26] Yates, F. (1960), Sample Methods for Censuses and Surveys. Charles Griffin and Sons, London, Third Edition.

- [14] Holt, Mary M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, rapport non publié, Research Triangle Institute, Research Triangle Park, North Carolina.
- [15] Kish, L., et Frankel, M.R. (1974), Inference from Complex Samples. Journal of the Royal Statistical Society B. 36, 1-37.
- [16] Konijn, H.S. (1962), Regression Analysis in Sample Surveys. Journal of the American Statistical Association 57, 590-606.
- [17] Kopitze, R., Boardman, T.J. et Graybill, F.A. (1975), Least Square Programs - A Look at the Square Root Procedure. The American Statistician, 29, 64-66.
- [18] Ling, R.G. (1974), Comparison of Several Algorithms for Computing Sample Means and Variances. Journal of the American Statistical Association, 69, 859-866.
- [19] Longley, James (1967), An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. Journal of the American Statistical Association, 62, 819-841.
- [20] Nathan, G. (1972), On the Asymptotic Power for Tests for Independence in Contingency Tables from Stratified Samples. Journal of the American Statistical Association, 67, 917-920.
- [21] Rao, J.N.K. (1975), Analytic Studies of Sample Survey Data. Techniques d'enquête, Statistique Canada.
- [22] Rao, J.N.K. (1975), Unbiased Variance Estimation for Multistage Designs, Sankhya C, 37, 133-139.

- [5] Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, 37, 117-132.
- [6] Fuller, W.A. et Hidiroglou, M.A. (1968), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.
- [7] Fuller, W.A. (1980a), Properties of Some Estimators for the Errors-in-Variables Model, Annals of Statistics, 8, 407-422.
- [8] Fuller, W.A. (1980b), Estimation of Measurement Error Models from Cluster Samples, Communication présentée aux réunions de l'American Educational Research Association. Boston, Massachusetts.
- [9] Garza-Hernandez, T. et McCarthy, P.J. (1962), A Test of Homogeneity for a Stratified Sample. Proceedings of the Social Statistics Section of the American Statistical Association, 200-202.
- [10] Gray, G.B. (1975), Components of Variance Model in Multistage Stratified Samples, Techniques d'enquête, Statistique Canada, 1, 27-43.
- [11] Hartley, H.O. (1959), Analytic Studies of Survey Data. Instituto de Statistica, Rome, Volume in onore di Corrado Gini.
- [12] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D., (1980), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.
- [13] Hidiroglou, M.A., Fuller, W.A., et Hickman, R.D. (1980), MINI CARP, Survey Section, Iowa State University, Ames, Iowa.



servir à repérer des erreurs dans la version indiquée et à les corriger. De plus, divers messages sont également imprimés. Par exemple, certains messages concernent les contrôles des données à l'entrée, comme lorsqu'on tente d'entrer plus de variables que le programme ne le permet, quand on tente d'introduire trop de grappes ou quand le format des données d'entrée ne répond pas aux normes établies. Si certaines strates contiennent une seule grappe, une liste de ces strates sera dressée. Dans le cas où l'utilisateur demande la combinaison de strates formées d'une grappe seulement, les strates ainsi produites sont imprimées.

SUPER CARP et MINI CARP sont écrits en FORTRAN avec double précision. Ils peuvent être utilisés sur les ordinateurs dotés d'un compilateur FORTRAN moyennant quelques petites modifications du langage de contrôle des travaux. On peut aussi ajouter de nouvelles techniques statistiques à SUPER CARP ou à MINI CARP. Ces techniques peuvent s'intégrer au programme sous la forme de sous-programmes qui peuvent être reliés aux autres éléments du logiciel.

## 6. BIBLIOGRAPHIE

- [1] Beaton, A.E., Rubin, D.B., et Barone, J.L. (1976), The Acceptability of Regression Solutions: Another Look at Computational Accuracy. Journal of the American Statistical Association, 71, 158-168.
- [2] Raj, Des (1968), Sampling Theory, McGraw-Hill Inc.
- [3] Raj, D. (1966), Some Remarks on a Simple Procedure of Sampling Without Replacement, Journal of the American Statistical Association, 61, 391-397.
- [4] Francis, I., Heiberg, R.M., et Veillemann, P.F. (1975), Criteria and Considerations in the Evaluation of Statistical Program Packages. The American Statistician, 29, 52-56.

Essentiellement, la documentation des utilisateurs est un guide qui explique à l'utilisateur la façon de se servir du programme. SUPER CARP et MINI CARP comprennent tous les deux un guide de ce genre qui se présente sous la forme suivante. D'abord, une introduction résume les diverses options statistiques offertes par la méthode. Ensuite, les données d'entrée et les commandes relatives aux analyses, aux variables et aux options sont expliquées, avec exemples à l'appui. Comme les données doivent être introduites et les commandes placées selon un ordre particulier, un organigramme d'analyse est inclus. Les techniques offertes par les programmes sont décrites en fonction des formules et des méthodes numériques utilisées, et avec quelques références à divers ouvrages.

Comme il a été mentionné précédemment, le langage de commande de SUPER CARP et de MINI CARP est constitué de codes numériques ou alphanumériques enregistrés dans des colonnes fixes sur des cartes. Francis, Heiberger et Velleman (1975) signalent que les langages de commande qui peuvent exécuter les calculs les plus efficaces sont fondés sur l'utilisation de codes numériques dans des colonnes fixes. L'inconvénient de cette méthode est que les utilisateurs doivent se référer trop souvent au manuel pour trouver des commandes. Il serait peut-être possible de permettre aux utilisateurs de demander des analyses et des options par des commandes semblables avec des mots anglais, en ajoutant un programme de traduction des commandes. L'avantage de cette méthode est qu'elle serait assez facile à apprendre, par contre le temps et le travail nécessaires pour programmer le traducteur en rendraient le coût prohibitif.

Les listes imprimées par SUPER CARP et MINI CARP indiquent la technique statistique appliquée et les variables utilisées dans l'analyse. Une partie de chaque liste imprimée montre le numéro de la version de SUPER CARP ou de MINI CARP et la date de la dernière mise à jour. Ces renseignements peuvent

formée d'une seule grappe. Pour s'assurer que cette exigence est respectée, il suffit d'enregistrer tous les groupes d'observations ayant une fraction d'échantillonnage égale à 1 au début du fichier. Quand deux strates sont regroupées, la fraction d'échantillonnage de la nouvelle strate est calculée en fonction de la fraction d'échantillonnage de chacune des deux strates combinées et du nombre d'éléments qu'elles contiennent.

#### 4.4 Grappes formées d'un seul élément

Si des grappes qui ne contiennent qu'un élément se trouvent dans une strate au premier degré d'échantillonnage, la combinaison de strates voisines permet de calculer des estimations de la variance. Dans un plan de sondage à plusieurs degrés, il peut arriver que certains degrés d'échantillonnage produisent des grappes formées d'un seul élément. Pour ce genre de grappes, il est impossible de calculer la variation à l'intérieur de l'ensemble des grappes pour lesquelles cette variation peut être calculée. L'incidence de ces degrés d'échantillonnage sur la variance, lorsque certaines grappes n'ont qu'un élément, peut alors être représentée par cette approximation.

### 5. QUELQUES CARACTÉRISTIQUES SOUHAITABLES DANS UN PROGRAMME D'ESTIMATION DE LA VARIANCE

Francis, Heiberg et Velleman (1975) ont dressé une liste de critères utiles d'évaluation des programmes statistiques en général. Dans la présente section, nous énumérons les caractéristiques qu'un programme informatique doit posséder pour l'estimation de la variance dans les enquêtes complexes. Parmi ces facteurs nécessaires, mentionnons la documentation des utilisateurs, les contrôles des données à l'entrée, les listes imprimées et l'efficacité statistique. Nous examinons ici dans quelle mesure les caractéristiques de SUPER CARP et MINI CARP répondent à ces besoins.

Lorsqu'une population échantillonnée est très hétérogène et qu'on applique plusieurs critères de stratification, il est fort possible que certaines strates contiennent seulement une grappe. Dans ce cas, il est impossible d'estimer la variabilité, et l'utilisateur peut demander que les strates composées d'une seule grappe soient combinées avec des strates voisines. Si cette demande n'est pas faite, SUPER CARP et MINI CARP excluent ces strates des calculs de la variance, mais les incluent pour les besoins d'estimation. Le programme produit une liste des strates à une seule grappe, ce qui peut aider l'utilisateur à combiner ce genre de strates quand il présente un programme par la suite. Pour cette combinaison, les strates à une grappe doivent avoir les mêmes caractéristiques que des strates voisines. On peut suggérer la méthode suivante qui se prête bien à la programmation. Lorsque le programme découvre une strate qui ne contient qu'une grappe, cette strate est fondue avec la strate suivante classée dans le fichier. Si la dernière strate est composée d'une seule grappe, la dernière strate est combinée avec l'avant-dernière. Une strate dont la fraction d'échantillonnage est égale à 1 n'est pas combinée parce qu'une telle strate n'influe pas sur la variance observée entre les unités primaires de l'échantillon. Les strates dont la fraction d'échantillonnage est égale à 1 ne doivent jamais figurer après une strate

#### 4.3 Combinaison de strates

L'analyse de Wilkinson (1965). De plus, cette technique permet de trouver les valeurs propres de systèmes d'équations ayant la forme  $Ax = Bx$ , où A est une matrice positive et B, une matrice semi-définie positive. Le calcul de valeurs propres est nécessaire dans SUPER CARP pour quelques-unes des analyses de régression avec erreurs sur les variables. Pour cette raison et compte tenu des critères de l'exactitude établis, on a adopté la décomposition de Cholesky comme méthode d'inversion dans SUPER CARP.



née à la préoccupation de s'assurer que les données sont assez précises pour que les résultats soient significatifs. Divers modèles d'algorithmes à un et à deux passages ont été étudiés par Ling (1974), qui est arrivé à la conclusion qu'il n'y a pas d'algorithme supérieur aux autres dans tous les cas. Le meilleur algorithme pour un ensemble de données en particulier dépend des chiffres contenus dans cet ensemble. Une des suggestions formulées par Ling est d'effectuer des calculs en double précision pour obtenir des résultats plus précis que ceux des calculs en simple précision. Les algorithmes récurrents à un passage sont préférable aux méthodes habituelles à un passage du type "machine de bureau", parce qu'ils tendent à produire moins d'erreurs de calcul. Cela se note surtout dans les sous-programmes avec simple précision. Dans SUPER CARP et MINI CARP, on a choisi des algorithmes récurrents à un passage programmés avec double précision.

#### 4.2 Inversion de matrices

L'inversion de matrices est nécessaire à la régression et à l'analyse de tableaux de contingence. Le choix de l'algorithme d'inversion dans une méthode informatique est très important, comme le démontre l'étude de Longley (1967), où l'auteur examine la précision de divers algorithmes d'inversion et découvre de sérieuses imperfections dans les calculs. Longley affirme avoir obtenu les résultats les plus précis en utilisant la technique d'orthonormalisation. Koptze, Boardman et Graybill (1975) préconisent l'application de la décomposition de Cholesky comme algorithme d'inversion. Ces auteurs mentionnent que, contrairement à la méthode d'élimination de Gauss, la technique de Cholesky ne requiert pas de pivot pour stabiliser les matrices définies positives symétriques. L'inversion prend donc moins de temps. La décomposition ne demande pas beaucoup de place en mémoire et elle est plus facile à programmer que la méthode d'élimination de Gauss. Un autre avantage de la décomposition de Cholesky est qu'elle est assez précise, comme le démontre



de l'information préliminaire d'ordre général, comme la définition du problème, le nombre d'observations à stocker, le support sur lequel se trouve les entrées (bande, disque ou cartes), la structure des données, de même que des renseignements sur la sortie des données et des contrôles pour la combinaison des strates. La carte du format indique la composition des données d'entrée ainsi que leur nature et le poids correspondant. La carte des noms des variables attribue des noms choisis aux zones des données d'entrée, suivant l'ordre dans lequel les données sont introduites. La carte de sélection contient les limites valables pour certaines variables lorsqu'une telle opération est nécessaire. La carte d'analyse énumère les analyses à effectuer (voir le tableau 1). Enfin, la carte d'identification des variables désigne les variables qui seront utilisées dans les analyses demandées. Parmi les cartes facultatives, on retrouve les cartes des fractions de sondage (on peut indiquer les pas de sondage de chaque strate), les cartes des erreurs sur les variables, qui donnent au programme la matrice des covariances pour les variables mesurées avec une erreur, et la carte des tests d'hypothèses qui permet de spécifier des coefficients de régression et de vérifier s'ils sont égaux à zéro.

#### 4. CALCULS

##### 4.1 Moyennes, sommes des carrés corrigées et produits vectoriels

Les moyennes, les sommes des carrés corrigées et les produits vectoriels sont des fonctions normalement traitées dans un programme d'enquête. Pour choisir les algorithmes nécessaires au calcul de ces fonctions, il faut prendre en considération le degré d'exactitude visé, la vitesse d'exécution et les contraintes liées au stockage des données. Beaton, Rubin et Barone (1976) ont admis que la recherche de méthodes de calcul très exactes doit être subordon-

Si nous définissons  $P_{ij|i}$  comme étant égal à  $P_{i+}^{-1} P_{ij}$  et désignons les estimations obtenues pour l'échantillon par la notation  $\hat{P}_{ij|i} = \hat{P}_{i+}^{-1} \hat{P}_{ij}$ , on peut estimer  $(P_{+1}, P_{+2}, \dots, P_{+,C-1})$  en effectuant la régression de  $\hat{P}_{ij|i}$  ( $i = 1, 2, \dots, R; j = 1, 2, \dots, C-1$ ) par rapport à des vecteurs-lignes à  $(C-1)$  dimensions qui contiennent 1 à la j<sup>ème</sup> entrée correspondant à  $\hat{P}_{ij|i}$  et des zéros ailleurs. Cette régression relève de la méthode des moindres carrés généralisée parce que la structure des erreurs des  $\hat{P}_{ij|i}$  n'est pas uniforme. On obtient une estimation de la matrice des covariances des  $\hat{P}_{ij|i}$ , qui tient compte du plan d'enquête, au moyen des formules de covariances élaborées pour la méthode des quotients. La variable utilisée dans le test de  $H_0$  est ensuite calculée à partir des sommes des carrés des résidus produits par la régression.

### 3. ENTREES

Dans une enquête, les données relatives à chaque unité choisie sont normalement caractérisées par la strate, le degré d'échantillonnage (premier, deuxième, ..., s<sup>ième</sup>) et un facteur de pondération. Les données doivent être classées selon cet ordre afin de produire des estimations de la variance qui correspondent à la structure des strates et des grappes.

SUPER CARP et MINI CARP utilisent un langage de commande composé de codes numériques placés à des positions fixes sur les cartes informatiques. Dans ces deux programmes, il y a six cartes de contrôle obligatoires qui doivent toujours faire partie des entrées. Un certain nombre de cartes de contrôle facultatives peuvent aussi être incluses lorsque des renseignements supplémentaires sont requis pour des options spécifiées sur les cartes obligatoires. Les cartes obligatoires comprennent la carte des paramètres, la carte des noms des variables, la carte du format, la carte de sélection, la carte d'analyse et la carte d'identification des variables. La carte des paramètres fournit

où  $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_{k-1})^T$  est le vecteur des proportions estimées à partir de la configuration des strates et des grappes,

$\hat{p}_0 = (p_{01}, p_{02}, \dots, p_{0,k-1})^T$  est le vecteur des proportions hypothétiques

$\hat{V} =$  la matrice des covariances de  $\hat{p}$  calculée à partir de la configuration des strates et des grappes,

$k =$  le nombre de catégories examinées,

$$d = \frac{1}{L} \sum_{h=1}^L (n_h - 1),$$

$L$  correspond au nombre de strates dans l'échantillon et  $n_i$  représente le nombre de grappes dans la  $i$ ème strate. On obtient la matrice des covariances  $\hat{V}$  par les méthodes décrites pour l'estimation par quotient. Dans les grands échantillons,  $F$  est distribuée approximativement comme une variable  $F$  centrée à  $k-1$  et  $d-k+2$  degrés de liberté lorsque l'hypothèse nulle est vraie.

Quant au test d'indépendance, Fuller (SUPER CARP, p. 65-69) a élaboré une technique qui tient compte du plan de sondage. Un tableau de contingence où la population est répartie selon deux critères contient  $R$  rangs et  $C$  colonnes, et l'hypothèse nulle qu'il faut vérifier se définit comme suit:  $H_0: p_{ij} = p_{i+} p_{+j}$  ou  $p_{ij} = p_{-i} p_{+j}$  où  $p_{ij}$  est la proportion de la population qui figure dans la  $(ij)$ ème case,  $p_{+j} = \sum_i p_{ij}$  et  $p_{i+} = \sum_j p_{ij}$ .

par rapport à celle des échantillons superposés répétés et équilibrés (balanced repeated replication) et à la technique dite "jack-knife". Premièrement, elle est assez facile à programmer et on peut l'adapter aux plans d'échantillonnage à plusieurs degrés. Ensuite, aucune contrainte n'est imposée au plan d'enquête (par exemple, deux échantillons répétés par strate) et les hypothèses sur lesquelles repose la méthode exigent un comportement régulier dans certains moments de la distribution de la population d'intérêt. Troisièmement, cette méthode requiert moins de calculs que les autres techniques.

Les données contiennent assez souvent des erreurs de mesure. Des notions théoriques pour les modèles de régression qui prennent en considération ces erreurs ont été présentées par Fuller (1980a), Fuller (1980b) et par Fuller et Hidiroglou (1978). SUPER CARP est également assez souple pour effectuer des tests d'hypothèses concernant n'importe quel sous-ensemble de paramètres de régression.

#### iv) Tableaux de contingence

SUPER CARP et MINI CARP exécutent le test de validité de l'ajustement ainsi que le test d'indépendance avec des données d'enquêtes complexes. Ces deux tests tiennent compte de la division en strates et en grappes prévue par le plan de sondage. Comme l'ont fait remarquer Rao et Scott (1981), les utilisateurs de la méthode classique du khi-carré de Pearson peuvent être sérieux-ment induits en erreur si les effets du plan d'enquête sont importants.

Pour le test de validité de l'ajustement, SUPER CARP et MINI CARP utilisent la valeur corrigée de la fonction discriminante de Wald, calculée à l'aide de la

formule suivante:

$$F_{WG} = [(k-1)d]^{-1} (d-k+2) (\hat{p}-p_0)^T \hat{V}^{-1} (\hat{p}-p_0)$$





### iii) Estimation par régression

On a accordé récemment beaucoup d'attention aux notions de régression qui s'appliquent aux enquêtes par sondage. Cet intérêt est attribuable à plusieurs facteurs. D'abord, on met davantage l'accent sur les enquêtes analytiques, bien que certaines questions concernant les meilleures méthodes de pondération des observations ne soient pas encore résolues. Deuxièmement, la construction de modèles en général, surtout ceux fondés sur les méthodes de régression, a suscité beaucoup d'intérêt, ainsi que des critiques, quant aux possibilités qu'elle présente pour le calcul des estimations d'enquête. SUPER CARP attribue le poids approprié aux observations et calcule la variance des estimations des coefficients de régression selon une méthode conçue par Fuller (1975).

Les coefficients de régression estimés à partir d'un échantillon stratifié en grappes proviennent de l'équation:

$$\tilde{b} = (X'WX)^{-1}X'WY$$

où le (rs)ième élément de  $(X'WX)$  est

$$\sum_{h=1}^L \sum_{i=1}^N x_{hi}^2 = \sum_{h=1}^L \sum_{i=1}^N x_{hi1}^2 + \sum_{h=1}^L \sum_{i=1}^N x_{hi2}^2 + \dots + \sum_{h=1}^L \sum_{i=1}^N x_{hij}^2$$

et le (r)ième élément de  $X'WY$  est

$$\sum_{h=1}^L \sum_{i=1}^N x_{hi} y_{hi} = \sum_{h=1}^L \sum_{i=1}^N x_{hi1} y_{hi1} + \sum_{h=1}^L \sum_{i=1}^N x_{hi2} y_{hi2} + \dots + \sum_{h=1}^L \sum_{i=1}^N x_{hij} y_{hij}$$

également des estimations sans biais pour  $\hat{V}(D_{\hat{Y}})$  et  $\hat{V}(D_{\hat{Y}})$ . Il devient alors possible d'appliquer les formules classiques pour  $\hat{V}$  et  $\hat{V}(\hat{Y})$  aux variables "synthétiques"  $D_{\hat{Y}}^{h_{js}}$ . On peut calculer des totaux pour chaque strate en considérant les strates comme des variables de classification.

#### ii) Estimation par quotient

Le vecteur  $\{Y_{h_{js}}(1), X_{h_{js}}(1), \dots, Y_{h_{js}}(p), X_{h_{js}}(p)\}$  est utilisé dans l'analyse et les rapports estimés sont les suivants:

$$\hat{R}(t) = \hat{X}_{-1}^{-1}(t) Y(t), \quad t = 1, 2, \dots, p;$$

où  $Y(1)$  et  $X(t)$  ont la forme présentée dans la section précédente. La matrice des covariances estimées de  $R = R(1), R(2), \dots, R(p)$  est la même que la matrice décrite dans la section précédente avec

$$\hat{D}_{h_{js}}^{h_{js}}(t) = \hat{X}_{-1}^{-1}(t) \begin{pmatrix} 1 & 2 & \dots & p \end{pmatrix} \begin{pmatrix} Y_{h_{js}}(1) \\ X_{h_{js}}(1) \\ \vdots \\ Y_{h_{js}}(p) \\ X_{h_{js}}(p) \end{pmatrix}; \quad t = 1, \dots, p.$$

On peut appliquer l'estimation par quotient au calcul de la moyenne de chaque variable d'intérêt en fixant toutes les variables  $X$  égales à 1. Il est possible de calculer la moyenne pour chaque domaine en substituant  $D_{h_{js}}^{h_{js}}(t)$  à  $Y_{h_{js}}(t)$  et  $D_{h_{js}}^{h_{js}}(t)$  à  $X_{h_{js}}(t)$ . Lorsqu'il faut obtenir les proportions de  $Y$  pour une sous-population et pour un domaine  $D_d$ , le numérateur du quotient est la somme pondérée des  $D_{h_{js}}^{h_{js}}(t)$  et le dénominateur est la somme pondérée des  $D_{h_{js}}^{h_{js}}(t)$ . Le rapport estimé entre deux variables définies dans un domaine  $D_d$  peut être évalué de la même façon. On peut également calculer des proportions et des rapports pour des strates qui représentent des variables de classification.

où  $\pi_{h12}$  représente la probabilité composée de sélection des unités choisies et 2. Lorsque  $n_h \geq 2$  et les probabilités composées de sélection ne sont pas connues, il est possible d'utiliser l'approximation de la variance pour l'échantillonnage sans remise formulée par Gray (1975). Gray démontre que les variances d'un échantillon constitué sans remise et avec des probabilités inégales de sélection peuvent être décomposées en une variance "avec remise" multipliée par un facteur de correction fini pour la population, ce facteur étant déterminé par les probabilités composées. Le facteur de correction a été évalué comme à peu près égal à 1 moins l'inverse de la fraction de sondage pour les populations composées de plus de 15 éléments à chaque degré. Au moyen de l'approximation de Gray, on peut calculer les variances pour les plans d'échantillonnage sans remise à plusieurs degrés avec des probabilités inégales de sélection.

S'il faut faire de l'estimation par domaine pour quelques variables, on définit une nouvelle variable,  $y_{hs}(k)$ , pour tous les éléments de la population, où

$y_{hs}(k)$  si le  $h^{\text{ième}}$  élément appartient au domaine

$d$  (disons  $D_d$ )

$\{ y_{hs}(k) = \begin{cases} 1 & \text{si le } h^{\text{ième}} \text{ élément appartient à } D_d \\ 0 & \text{autrement} \end{cases}$

Une autre façon de définir  $y_{hs}(k)$  consiste à appliquer la formule  $y_{hs}(k) = d y_{hs}(k)$ , où

$d y_{hs}(k) = \begin{cases} 1 & \text{si le } h^{\text{ième}} \text{ élément appartient à } D_d \\ 0 & \text{autrement} \end{cases}$

Notons que, si  $\hat{y}$  et  $v(\hat{y})$  sont des estimateurs sans biais pour  $y$  et  $v(y)$ , respectivement, les résultats de l'estimation par domaine,  $\hat{d y}$  et  $v(d y)$ , sont

primaires d'échantillonnage initiales comme des "strates" et les unités secondaires sont lues la  $r$ ème fois, les unités initialement du  $(r-1)$ ème degré d'échantillonnage sont considérées comme des "strates" et les unités du  $r$ ème degré comme des grappes, ce qui donne  $v_2(\hat{X})$ . Enfin, lorsque les  $(r \geq 2)$  données sont lues la  $r$ ème fois, les unités initialement du  $(r-1)$ ème degré d'échantillonnage sont considérées comme des "strates" et les unités du  $r$ ème degré comme des grappes, pour calculer  $v_r(\hat{X})$ .

À chaque passage en machine des données, une fraction d'échantillonnage  $g_{h_i r-1}$  est enregistrée pour l'unité  $h_i r-1$ , où

$$g_{h_i r-1} = 1 - \left[ \frac{n_{h_i j}}{N_{h_i j}} \right]_{j=0}^{r-2} - \left[ 1 - \frac{n_{h_i r-1}}{N_{h_i r-1}} \right] .$$

Par cette méthode, le programme calcule  $v_r(\hat{X})$  selon la structure donnée par  $v_1(\hat{X})$ .

Si les fractions de sondage ne sont pas négligeables à chaque degré et si l'enquête est basée sur un plan d'échantillonnage sans remise avec probabilité proportionnelle à la taille à tous les degrés, il faut prendre en considération la probabilité composée de sélection dans l'estimation de la variance à chaque degré. SUPER CARP et MINI CARP ne calculent pas les probabilités composées de sélection. Lorsque deux unités par strate ont été sélectionnées sans remise avec des probabilités inégales, on peut obtenir la variance de l'estimateur du total à l'aide de l'équation (2.2.1) en incluant, pour chaque strate, un facteur de correction qui contient la probabilité composée de sélection. Ce facteur de correction est le résultat de l'expression suivante:

$$f_h = \frac{2 \pi_{h12} - \pi_{h1} \pi_{h2}}{\pi_{h12}} , \quad h=1, 2, \dots, L$$

$$v_I(\tilde{X}) = \sum_{h=1}^L \sum_{i=1}^{n_h} \dots \sum_{i=1}^{n_{I-1}} \left[ \frac{n_{h_j}}{n_{j=0}} \right]$$

(2.2.2)

$$x \cdot n_{h_{I-1}}^{I-1} (n_{h_{I-1}}^{I-1} - 1) \dots (1 - f_{h_{I-1}}^{I-1})$$

$$x \cdot \sum_{i=1}^I (d_{h_{I-1}}^{I-1}(\cdot) - \tilde{d}_{h_{I-1}}^{I-1}(\cdot))$$

$$x \cdot d_{h_{I-1}}^{I-1}(\cdot) - \tilde{d}_{h_{I-1}}^{I-1}(\cdot)$$

ou

$$f_{h_{I-1}}^{I-1} = n_{h_{I-1}}^{I-1}, n_{h_{I-1}}^{I-1-1}, n_{h_{I-1}}^{I-1} = n, n_{h_{I-1}}^{I-1} = N, n_{h_{I-1}}^{I-1}$$

$$d_{h_{I-1}}^{I-1}(\cdot) = d_{h_{I-1}}^{I-1}(1), d_{h_{I-1}}^{I-1}(2), \dots, d_{h_{I-1}}^{I-1}(p)$$

$$d_{h_{I-1}}^{I-1}(k) = \sum_{i=1}^{I-1} s_{h_{I-1}}^i \dots s_{h_{I-1}}^i X_{h_{I-1}}(k)$$

$$\frac{d}{d} d_{h_{I-1}}^{I-1}(\cdot) = \sum_{i=1}^I d_{h_{I-1}}^{I-1}(\cdot)$$

On peut donc estimer la variance pour un plan de sondage à  $r$  degrés en calculant les composantes qui correspondent à chaque degré d'échantillonnage ( $v_I(\tilde{X})$ ) et en faisant l'addition. On peut faire ce calcul en passant  $r$  fois en machine l'ensemble de données. La première fois, le programme repère les strates et les unités du premier degré d'échantillonnage pour obtenir  $v_I(\tilde{X})$ . La deuxième fois, le programme enregistre les unités



$$\tilde{X} = \sum_{s=1}^s \tilde{X}_s$$

où pour  $i \geq 2$

Il est à noter que la formule de variance ci-dessus peut être appliquée à des plans d'échantillonnage avec probabilité proportionnelle à la taille (ppt) avec ou sans remise. Dans le cas de l'échantillonnage avec remise, il faut calculer seulement la variance au premier degré (Des Raj, 1968, p. 120) et les facteurs de correction,  $f_h$ , sont fixés à zéro. Dans les enquêtes à grande échelle, il est souvent supposé que les grappes du premier degré d'échantillonnage ont été sélectionnées avec remise, même si, en réalité, l'échantillonnage était fait sans remise. Un corollaire de cette hypothèse et du fait que les pas de sondage sont petits est que la variance ainsi calculée se rapproche beaucoup de celle qu'on aurait obtenue en tenant compte de tous les degrés d'échantillonnage et de toutes les techniques de sélection. Si les pas de sondage sont assez importants à chaque degré d'échantillonnage et si l'échantillon a été prélevé par échantillonnage aléatoire simple sans remise à chaque degré, on peut tirer avantage de la règle de Des Raj (1966) pour calculer la composante de la variance à chaque degré d'échantillonnage. La matrice des covariances pour un échantillon formé de  $s$  degrés s'écrit :

$$D_{h_1 h_2} = \sum_{i=1}^I D_{h_1 h_2}^{(i)} \quad \dots \quad \sum_{i=1}^I D_{h_1 h_2}^{(i)} = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K D_{h_1 h_2}^{(i,j,k)}$$

où Y représente la variable dépendante et X, les variables indépendantes s'il s'agit d'une analyse de régression. Il est à noter que  $v_1$  correspond toujours à la variable dépendante dans le cas d'une régression. Quant aux autres types d'analyse, l'ordre à l'intérieur du vecteur de sélection n'a pas d'importance.

## 2.2 Types de calculs

Cette section présente un aperçu des variables statistiques simples définies dans les programmes et donne une liste partielle des options de régression offertes. Une description complète de toutes les options est donnée dans le guide de SUPER CARP ou de MINI CARP (1980).

### 1) Estimateur de totaux

L'estimateur de totaux se formule comme suit:

$$\hat{X}_{(k)} = \sum_{i=1}^n \sum_{j=1}^p w_{ij} x_{ij}(k), \quad k = 1, 1, 2, \dots, p.$$

La matrice des covariances estimées de

$$\hat{X} = \{ \hat{X}_{(1)}, \hat{X}_{(2)}, \dots, \hat{X}_{(p)} \}$$

est

$$V(\hat{X}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(1 - f_i) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n} \sum_{i=1}^n x_i \bar{x} - \frac{1}{n} \sum_{i=1}^n \bar{x} x_i + \frac{1}{n} \sum_{i=1}^n \bar{x}^2$$

où

$$\hat{d}_{h1} = \{ \hat{d}_{h1}(1), \hat{d}_{h1}(2), \dots, \hat{d}_{h1}(p) \}$$

## 2. DESCRIPTION GÉNÉRALE

### 2.1 Notation

En général, SUPER CARP et MINI CARP peuvent être utilisés avec des données obtenues à l'aide d'un plan stratifié à plusieurs degrés. Si le plan de sondage comporte  $s$  degrés d'échantillonnage, un vecteur de degrés à  $g$  dimensions est mis en mémoire pour chaque observation. On peut représenter ce vecteur de données de la façon suivante:

$$(z_{h1}^{s1}, z_{h2}^{s2}, \dots, z_{hg}^{sg})$$

où  $h = 1, 2, \dots, L$  représente les strates,  $is = (i1, i2, \dots, is)$  indique le degré d'échantillonnage,  $i1 = 1, 2, \dots, nh$  correspond aux unités du premier degré d'échantillonnage,  $i2 = 1, 2, \dots, nh11$  correspond aux unités du deuxième degré d'échantillonnage,  $\dots$ , et  $is = 1, 2, \dots, nh1$  désigne les unités du dernier degré d'échantillonnage,  $s$ .  $zh1k$  représente la  $h$ ème observation pour la  $k$ ème variable d'intérêt. Un poids attribué à la  $h$ ème observation sera désigné par le terme  $wh1$ . Ces poids sont inversement proportionnels à la probabilité que chaque unité finale d'échantillonnage soit sélectionnée. On énumère les variables qui serviront à l'analyse (que ce soit un total, une estimation par quotient ou par régression) au moyen d'un vecteur de sélection  $v = (v1, v2, \dots, vp+1)$ , où  $1 \leq vk \leq g$  pour  $k = 1, 2, \dots, p+1$ . Supposons que le type d'analyse et les variables requises ont été choisis par l'utilisateur, et disons que le vecteur de sélection de la  $h$ ème observation est le suivant:

$$(y_{h1}^{s1}, x_{h1}^{s1}, x_{h2}^{s2}, \dots, x_{hg}^{sg})$$

données auxquelles les programmes sont appliqués et imputation des données qui manquent. Dans les pages qui suivent, nous présentons une description de SUPER CARP et de MINI CARP. À l'aide de SUPER CARP, on peut calculer des estimations de totaux, des estimations par quotient, la différence entre des estimations obtenues par la méthode des quotients et construire des tests applicables à des tableaux de contingence pour des échantillons stratifiés à plusieurs degrés. SUPER CARP offre quelques techniques de régression qui conviennent à des observations touchées par des erreurs de réponse (de mesure). L'utilisateur peut estimer des matrices des covariances pour les moyennes et les totaux de sous-populations ou de strates. MINI CARP est un programme plus petit qui diffère de SUPER CARP en ce sens qu'il ne contient pas les techniques de régression de SUPER CARP. Une comparaison des applications de ces deux programmes est présentée au tableau 1.

TABLEAU 1. Applications de SUPER CARP (S) et de MINI CARP (M)

Estimation à plusieurs variables de	Pour		
	Toute la population	Des strates individuelles	Des sous- populations
Paramètres simples			
• Moyennes	S, M	S, M	S, M
• Totaux	S, M	S, M	S, M
• Quotients	S, M	S, M	S, M
• Différences entre quotients	S, M	S, M	S, M
• Proportions	S, M	S, M	S, M
Paramètres complexes			
• Moindres carrés pondérés	S		
• Erreurs pondérées sur les variables (covariances connues et estimées des erreurs)	S		
Tests			
• Coefficients de régression	S		
• Validité de l'ajustement	S, M		
• Indépendance, tableau à deux dimensions	S, M		



méthodes utilisées pour l'estimation par domaine. Les estimateurs de la variance des estimations obtenues pour les domaines ne sont que des modifications directes des estimateurs de la variance appliqués aux variables statistiques simples. Cela n'est toutefois pas le cas des variables très complexes. L'estimation d'équations de régression basées sur des données d'enquêtes pose plusieurs problèmes. En effet, il faut formuler les équations de régression, définir la population sur laquelle les inférences porteront et estimer la variance des coefficients de l'équation (voir Konijn (1962), Kish et Frankel (1974) et Fuller (1975)). Les tests d'hypothèses appliqués aux tableaux de contingence en tenant compte du plan de sondage ont été analysés par Nathan (1969, 1972), Rao et Scott (1981), Fellegi (1980), Garza-Hernandez et McCarthy (1962) et par Koch, Freeman et Freeman (1975), pour ne nommer que ces auteurs-là.

Très souvent, les organismes statistiques chargés de recueillir des données d'enquête s'acquittent aussi de toutes les étapes du traitement jusqu'à la mise en tableau des résultats. En outre, les programmes informatiques appliqués aux observations sont, dans la plupart des cas, adaptés au plan d'enquête. Il est fort possible que chaque fois qu'un nouveau plan de sondage est appliqué, il faille mettre au point des programmes tout à fait nouveaux pour, par exemple, estimer des totaux et leur variance. Ce travail prend beaucoup de temps, coûte très cher et il est fastidieux, parce que, dans une certaine mesure, répétitif. Une autre possibilité consiste à utiliser un ensemble de logiciels statistiques comme le SPSS ou le SAS. Ces programmes informatiques permettent d'obtenir facilement des estimations pondérées. Toutefois, les variances calculées par ces procédés ne prennent pas en considération des facteurs liés au plan de sondage, comme la stratification ou la division de l'échantillon en grappes, à moins que les instructions nécessaires ne soient ajoutées aux programmes. L'utilisateur doit donc connaître assez bien le langage de ces méthodes s'il veut obtenir de bonnes estimations de la variance des paramètres d'un sondage.

Des travaux ont été effectués récemment dans le but d'élaborer des programmes qui calculent des variances pour une catégorie générale de plans d'enquêtes. Parmi ces programmes, on trouve, entre autres, STDEER de Shah (1974), SURREGR de Holt (1975), ainsi que SUPER CARP et MINI CARP de Hidiroglou, Fuller et Hickman (1980). Essentiellement, ces programmes exigent que l'utilisateur indique l'estimateur qu'il veut utiliser et les variables qu'il doit analyser. Nous supposons ici qu'il y a vérification des ensembles de



## INFORMATISATION DU CALCUL D'ESTIMATIONS POUR LES ENQUÊTES COMPLEXES<sup>1</sup>

M.A. Hidyoglou<sup>2</sup>

Très souvent, les organismes statistiques chargés de recueillir des données d'enquête s'acquittent aussi de toutes les étapes du traitement jusqu'à la mise en tableau des résultats. En outre, les programmes informatiques appliqués aux observations sont, dans la plupart des cas, adaptés au plan d'enquête. Les résultats de ces calculs peuvent varier de paramètres descriptifs simples, comme un total ou une moyenne, aux paramètres plus complexes nécessaires aux études analytiques telles que la comparaison de domaines, l'analyse de régression et l'analyse de tableaux de contingence. Cet article décrit un programme informatique qui calcule ces statistiques et les erreurs d'échantillonnage correspondantes pour certains plan d'échantillonnage courants.

## 1. INTRODUCTION

Un grand nombre de statistiques sont calculées à partir de données qui, dans bien des cas, sont recueillies dans de grandes enquêtes complexes à l'échelle nationale ou régionale. Les résultats de ces calculs peuvent varier de paramètres descriptifs simples, comme un total ou une moyenne, aux paramètres plus complexes nécessaires aux études analytiques telles que la comparaison de domaines, l'analyse de régression et l'analyse de tableaux de contingence. L'estimation par domaine concerne la production de statistiques sur des sous-groupes de la population observée, qui ne sont pas explicitement prévus dans le plan de sondage. Yates (1960) offre une mine de renseignements sur l'estimation des moyennes et des écarts entre moyennes au niveau des domaines. Hartley (1959) et Rao (1975) donnent une excellente description des

<sup>1</sup> Exposé présenté à l'assemblée annuelle de l'American Statistical Association, Détroit, août 1981.  
<sup>2</sup> M.A. Hidyoglou Division des méthodes d'enquêtes "Entreprises", Statistique Canada

TABLEAU 5. Indice du biais de renouvellement par groupe d'âge et sexe (1981)

Caractéristique	Groupe d'âge-sexe	Nombre de mois d'inclusion dans l'échantillon					
		1	2	3	4	5	6
Personnes occupées	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
	F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
	M 25+	97.0	101.4	101.3	100.7	100.1	99.5
	F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
	F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
Chômeurs	M 25+	98.0	102.1	101.6	100.3	98.0	100.1
	F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
	F 15-24	96.8	100.0	100.7	100.7	101.0	100.8
	M 25+	97.1	101.4	101.3	100.7	100.0	99.5
	F 25+	97.1	101.3	101.4	101.0	100.1	99.1
Personnes actives	M 15-24	98.2	97.2	98.8	100.3	101.9	103.5
	F 15-24	97.6	97.2	99.3	100.3	102.1	103.5
	M 25+	98.7	98.9	99.4	100.1	101.0	101.8
	F 25+	98.6	98.8	99.3	100.3	101.3	101.6
	M 15-24	102.6	100.0	99.3	100.1	97.7	100.3
	F 15-24	104.2	100.3	96.0	98.3	100.8	100.4
Chômeurs	M 25+	99.6	99.8	99.8	99.7	98.8	102.3
	F 25+	101.8	100.0	102.6	99.8	96.1	99.6
	M 15-24	98.9	97.6	98.9	100.3	101.3	103.0
	F 15-24	98.5	97.6	98.8	100.0	101.9	103.2
	M 25+	98.8	99.0	99.5	100.1	100.8	101.0
	F 25+	98.8	98.8	99.6	100.3	101.0	101.5
Personnes occupées	M 15-24	99.9	99.0	99.7	100.0	100.5	101.0
	F 15-24	99.3	99.1	100.2	100.0	100.6	101.0
	M 25+	100.4	100.7	100.3	99.7	99.5	99.3
	F 25+	100.3	100.6	100.2	100.0	99.9	99.1
	M 15-24	104.2	101.7	100.1	99.7	96.3	98.0
	F 15-24	105.8	102.0	96.8	98.0	99.4	98.0
Chômeurs	M 25+	101.2	101.6	100.7	99.4	97.4	99.8
	F 25+	103.4	101.7	103.5	99.5	94.8	97.1
	M 15-24	100.5	99.4	99.8	99.9	99.8	100.5
	F 15-24	100.2	99.4	99.7	99.7	100.4	100.6
	M 25+	100.4	100.8	100.3	99.7	99.4	99.3
	F 25+	100.5	100.6	100.4	99.9	99.5	98.9
Personnes actives	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
	F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
	M 25+	97.0	101.4	101.3	100.7	100.1	99.5
	F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
	F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
Chômeurs	M 25+	98.0	102.1	101.6	100.3	98.0	100.1
	F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
	F 15-24	96.8	100.0	100.7	100.7	101.0	100.8
	M 25+	97.1	101.4	101.3	100.7	100.0	99.5
	F 25+	97.1	101.3	101.4	101.0	100.1	99.1
Personnes occupées	M 15-24	98.2	97.2	98.8	100.3	101.9	103.5
	F 15-24	97.6	97.2	99.3	100.3	102.1	103.5
	M 25+	98.7	98.9	99.4	100.1	101.0	101.8
	F 25+	98.6	98.8	99.3	100.3	101.3	101.6
	M 15-24	102.6	100.0	99.3	100.1	97.7	100.3
	F 15-24	104.2	100.3	96.0	98.3	100.8	100.4
Chômeurs	M 25+	99.6	99.8	99.8	99.7	98.8	102.3
	F 25+	101.8	100.0	102.6	99.8	96.1	99.6
	M 15-24	98.9	97.6	98.9	100.3	101.3	103.0
	F 15-24	98.5	97.6	98.8	100.0	101.9	103.2
	M 25+	98.8	99.0	99.5	100.1	100.8	101.0
	F 25+	98.8	98.8	99.6	100.3	101.0	101.5
Personnes occupées	M 15-24	99.9	99.0	99.7	100.0	100.5	101.0
	F 15-24	99.3	99.1	100.2	100.0	100.6	101.0
	M 25+	100.4	100.7	100.3	99.7	99.5	99.3
	F 25+	100.3	100.6	100.2	100.0	99.9	99.1
	M 15-24	104.2	101.7	100.1	99.7	96.3	98.0
	F 15-24	105.8	102.0	96.8	98.0	99.4	98.0
Chômeurs	M 25+	101.2	101.6	100.7	99.4	97.4	99.8
	F 25+	103.4	101.7	103.5	99.5	94.8	97.1
	M 15-24	100.5	99.4	99.8	99.9	99.8	100.5
	F 15-24	100.2	99.4	99.7	99.7	100.4	100.6
	M 25+	100.4	100.8	100.3	99.7	99.4	99.3
	F 25+	100.5	100.6	100.4	99.9	99.5	98.9
Personnes actives	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
	F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
	M 25+	97.0	101.4	101.3	100.7	100.1	99.5
	F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
	F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
Chômeurs	M 25+	98.0	102.1	101.6	100.3	98.0	100.1
	F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
	F 15-24	96.8	100.0	100.7	100.7	101.0	100.8
	M 25+	97.1	101.4	101.3	100.7	100.0	99.5
	F 25+	97.1	101.3	101.4	101.0	100.1	99.1
Personnes occupées	M 15-24	98.2	97.2	98.8	100.3	101.9	103.5
	F 15-24	97.6	97.2	99.3	100.3	102.1	103.5
	M 25+	98.7	98.9	99.4	100.1	101.0	101.8
	F 25+	98.6	98.8	99.3	100.3	101.3	101.6
	M 15-24	102.6	100.0	99.3	100.1	97.7	100.3
	F 15-24	104.2	100.3	96.0	98.3	100.8	100.4
Chômeurs	M 25+	99.6	99.8	99.8	99.7	98.8	102.3
	F 25+	101.8	100.0	102.6	99.8	96.1	99.6
	M 15-24	98.9	97.6	98.9	100.3	101.3	103.0
	F 15-24	98.5	97.6	98.8	100.0	101.9	103.2
	M 25+	98.8	99.0	99.5	100.1	100.8	101.0
	F 25+	98.8	98.8	99.6	100.3	101.0	101.5
Personnes occupées	M 15-24	99.9	99.0	99.7	100.0	100.5	101.0
	F 15-24	99.3	99.1	100.2	100.0	100.6	101.0
	M 25+	100.4	100.7	100.3	99.7	99.5	99.3
	F 25+	100.3	100.6	100.2	100.0	99.9	99.1
	M 15-24	104.2	101.7	100.1	99.7	96.3	98.0
	F 15-24	105.8	102.0	96.8	98.0	99.4	98.0
Chômeurs	M 25+	101.2	101.6	100.7	99.4	97.4	99.8
	F 25+	103.4	101.7	103.5	99.5	94.8	97.1
	M 15-24	100.5	99.4	99.8	99.9	99.8	100.5
	F 15-24	100.2	99.4	99.7	99.7	100.4	100.6
	M 25+	100.4	100.8	100.3	99.7	99.4	99.3
	F 25+	100.5	100.6	100.4	99.9	99.5	98.9
Personnes actives	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
	F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
	M 25+	97.0	101.4	101.3	100.7	100.1	99.5
	F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
	F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
Chômeurs	M 25+	98.0	102.1	101.6	100.3	98.0	100.1
	F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
	F 15-24	96.8	100.0	100.7	100.7	101.0	100.8
	M 25+	97.1	101.4	101.3	100.7	100.0	99.5
	F 25+	97.1	101.3	101.4	101.0	100.1	99.1
Personnes occupées	M 15-24	98.2	97.2	98.8	100.3	101.9	103.5
	F 15-24	97.6	97.2	99.3	100.3	102.1	103.5
	M 25+	98.7	98.9	99.4	100.1	101.0	101.8
	F 25+	98.6	98.8	99.3	100.3	101.3	101.6
	M 15-24	102.6	100.0	99.3	100.1	97.7	100.3
	F 15-24	104.2	100.3	96.0	98.3	100.8	100.4
Chômeurs	M 25+	99.6	99.8	99.8	99.7	98.8	102.3
	F 25+	101.8	100.0	102.6	99.8	96.1	99.6
	M 15-24	98.9	97.6	98.9	100.3	101.3	103.0
	F 15-24	98.5	97.6	98.8	100.0	101.9	103.2
	M 25+	98.8	99.0	99.5	100.1	100.8	101.0
	F 25+	98.8	98.8	99.6	100.3	101.0	101.5
Personnes occupées	M 15-24	99.9	99.0	99.7	100.0	100.5	101.0
	F 15-24	99.3	99.1	100.2	100.0	100.6	101.0
	M 25+	100.4	100.7	100.3	99.7	99.5	99.3
	F 25+	100.3	100.6	100.2	100.0	99.9	99.1
	M 15-24	104.2	101.7	100.1	99.7	96.3	98.0
	F 15-24	105.8	102.0	96.8	98.0	99.4	98.0
Chômeurs	M 25+	101.2	101.6	100.7	99.4	97.4	99.8
	F 25+	103.4	101.7	103.5	99.5	94.8	97.1
	M 15-24	100.5	99.4	99.8	99.9	99.8	100.5
	F 15-24	100.2	99.4	99.7	99.7	100.4	100.6
	M 25+	100.4	100.8	100.3	99.7	99.4	99.3
	F 25+	100.5	100.6	100.4	99.9	99.5	98.9
Personnes actives	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
	F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
	M 25+	97.0	101.4	101.3	100.7	100.1	99.5
	F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
	F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
Chômeurs	M 25+	98.0	102.1	101.6	100.3	98.0	100.1
	F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
	F 15-24	96.8	100.0	100.7	100.7	101.0	

TABLEAU 3. Indice du biais de renouvellement pour la population par type de région

Poids	Type de région	1	2	3	4	5	6
Non ajusté	AR	97.0	101.1	101.2	100.6	100.2	99.2
Ajusté pour les ménages	AR	98.7	98.7	99.4	100.0	101.1	101.0
Ajusté pour la population	AR	100.4	100.5	100.2	99.7	99.6	99.9
	UNAR	99.3	98.6	99.0	100.3	101.1	101.0
	UNAR	97.7	101.0	100.8	100.9	100.2	99.9
	UNAR	100.9	100.3	99.8	99.9	99.6	99.9

TABLEAU 4. Indice du biais de renouvellement par type de région (1981)

Poids	Caractéris- tique	Type de région	1	2	3	4	5
Non ajusté	Personnes occupées	AR	99.9	101.0	101.3	100.7	100.4
	Personnes occupées	UNAR	96.8	100.9	100.6	101.2	99.7
	Chômeurs	AR	99.1	102.6	101.3	100.4	97.7
	Chômeurs	UNAR	103.3	101.5	101.4	99.8	96.5
	Personnes actives	AR	97.0	101.1	101.3	100.7	100.2
	Personnes actives	UNAR	97.3	100.9	100.7	101.1	99.3
Ajusté pour les ménages	Personnes occupées	AR	98.6	98.5	99.5	100.1	101.2
	Personnes occupées	UNAR	98.3	98.4	98.7	100.6	101.6
	Chômeurs	AR	100.8	100.3	99.5	99.8	98.5
	Chômeurs	UNAR	104.9	99.2	99.6	99.2	97.3
	Personnes actives	AR	98.7	98.7	99.5	100.1	101.0
	Personnes actives	UNAR	98.9	98.4	98.8	100.5	101.2
Ajusté pour la population	Personnes occupées	AR	100.2	100.3	100.4	99.7	99.7
	Personnes occupées	UNAR	100.0	100.2	99.6	100.2	100.1
	Chômeurs	AR	102.4	102.1	100.4	99.4	97.1
	Chômeurs	UNAR	106.4	100.8	100.5	98.9	96.0
	Personnes actives	AR	100.4	100.5	100.4	99.7	99.5
	Personnes actives	UNAR	100.6	100.2	99.7	100.1	99.8

EAU 1. % Taux de non-réponse des ménages, en pourcentage, selon le nombre de mois d'inclusion dans l'échantillon et le type de région (1981)

Nombre de mois	UNAR	Type de région	AR	Canada†
----------------	------	----------------	----	---------

1	6.6	7.9	7.3	
2	4.0	4.6	4.4	
3	3.5	4.4	3.9	
4	3.5	4.1	3.8	
5	3.2	3.8	3.6	
6	3.1	3.6	3.4	

Le nombre moyen de ménages	26,707	28,645	55,352	L'exclusion des secteurs spéciaux
----------------------------	--------	--------	--------	-----------------------------------

EAU 2. Proportions estimées des chefs de ménage occupés et en chômage chez les ménages répondants et non répondants

Proportion	Répondants	Non-Répondants	Occupés	Chômeurs	Occupés	Chômeurs
	$\frac{Y_1}{Z_1}$	$\frac{Y_2}{Z_2}$	$\frac{Y_1}{Z_1} - \frac{Y_2}{Z_2}$	$\frac{Y_1}{Z_1}$	$\frac{Y_2}{Z_2}$	$\frac{Y_1}{Z_1} - \frac{Y_2}{Z_2}$

0.6893	0.0383	0.7839	0.0335	-0.0946	0.0048
0.6962	0.0344	0.7841	0.0321	-0.0879	0.0023
0.7006	0.0311	0.7851	0.0300	-0.0845	0.0011
0.7006	0.0364	0.7877	0.0281	-0.0871	0.0083
0.6972	0.0317	0.7821	0.0317	-0.0849	0.0000
0.6927	0.0331	0.7767	0.0320	-0.0840	0.0011
0.6961	0.0342	0.7833	0.0311	-0.0872	0.0031



## REMERCIEMENTS

L'auteur tient à remercier R. Vettore et R. Barnes pour l'élaboration de programmes informatiques, de même que les arbitres pour leurs remarques pertinentes.

## BIBLIOGRAPHIE

- [1] Bailar, B.A., (1975), The Effect of Rotation Group Bias on Estimates from Panel Surveys, JASA vol. 70, pp. 23-30.
- [2] Bailar, B.A., Bailey, L. et Corby, C., (1978), A Comparison of Some Adjustment and Weighting Procedures for Survey Data. Survey Sampling and Measurement, Academic Press, pp.175-198.
- [3] Platek, R., Singh, M.P. et Tremblay, V., (1978), Adjustment for Non-Response in Surveys, Techniques d'enquête, vol. 3, n° 1, pp. 1-24.
- [4] Statistique Canada (1976), Méthodologie de l'enquête sur la population active du Canada, n° 71-526 au catalogue, hors série.
- [5] Tessier, R. et Tremblay, V., (1976), Findings on Rotation Group Biases, document interne, Statistique Canada.
- [6] Thomsen, I., (1973), A Note on the efficiency of Sub-class Means to Reduce the Effects of Non-response when Analyzing Survey Data, Statistical Review, publié par le National Central Bureau of Statistics, Stockholm, Suède, vol. 11, n° 4.
- [7] Williams, W.H. et Mallows, C.L., (1970), Systematic Biases in Panel Surveys Due to Differential Non-Response, JASA, vol. 65, n° 331.
- [8] Woltman, H. et Bushery, J., (1975), A Panel Bias Study in the National Crime Survey, document présenté à la réunion annuelle de l'ASA.



en-dessous de la moyenne, au niveau des ménages, qui sont corrigés pour compenser les variations du nombre de personnes dénombrees d'un groupe de renouvellement à l'autre.

## 6. RÉSUMÉ ET CONCLUSION

La présente étude examine un modèle qui divise l'ensemble du biais en trois composantes, nommément l'effet des différences observées d'un groupe à l'autre de l'échantillon entre les taux de réponse, entre le biais de réponse et entre les caractéristiques des répondants et des non-répondants. Ce modèle indique aussi l'incidence de divers autres facteurs sur le biais de renouvellement. Il est possible de considérer les groupes de renouvellement comme un cas particulier de ce genre de groupes où l'ajustement du poids pour compenser la non-réponse peut être fait séparément.

Si les taux de réponse varient d'un groupe de renouvellement à l'autre et si la proportion de répondants et de non-répondants qui possèdent une caractéristique et le biais de réponse correspondant sont identiques pour tous les groupes de renouvellement, la correction de la non-réponse appliquée à chaque groupe de renouvellement ne change pas le biais des estimations. Toutefois, le biais de renouvellement peut augmenter ou diminuer, si le taux de réponse est respectivement en dessous ou au-dessus du taux de réponse moyen. Cette conclusion est confirmée par les données sur les valeurs de l'indice obtenues avant et après l'ajustement du poids en fonction des chiffres sur les personnes.

On entend faire des analyses des valeurs de l'indice pour des estimations relatives à l'activité sur le marché du travail et d'autres caractéristiques calculées à partir d'ensembles de données plus grands. On étudiera l'effet des différences entre la taille moyenne des ménages dans les groupes de renouvellement et parmi les ménages répondants et non répondants sur les estimations du biais de renouvellement. L'analyse portera également sur le rôle des différences entre les taux de réponse et des variations du biais de réponse, après la correction de la non-réponse, dans chaque groupe de renouvellement.

Les tableaux 4 et 5 présentent des données sur les valeurs moyennes de l'indice selon le type de région et le groupe d'âge et sexe, pour les douze enquêtes menées en 1981. Les valeurs obtenues par type de région sans ajustement du poids indiquent que le biais relatif des estimations concernant les chômeurs a tendance à être positif dans les deux premiers mois et à diminuer dans les derniers mois. Le biais relatif des estimations concernant les personnes occupées et les personnes actives a tendance à être négatif au premier mois et positif dans les mois subséquents. Les données sur les valeurs de l'indice calculées pour les groupes d'âge et les deux sexes affichent des tendances semblables à celles observées pour les types de région.

L'ajustement du poids pour compenser la non-réponse à l'aide des chiffres sur les ménages a tendance à augmenter les valeurs de l'indice au premier ainsi qu'au cinquième et au sixième mois. Les valeurs de l'indice obtenues pour les autres mois ont tendance à diminuer. Cette observation s'applique aux valeurs de l'indice calculées pour les catégories d'activité en fonction du type de région et du groupe d'âge et du sexe. On peut attribuer l'augmentation des valeurs de l'indice, au premier mois, à des taux de réponse en-dessous de la moyenne et la diminution des valeurs dans les mois subséquents, à des taux de réponse au-dessus de la moyenne. La baisse observée aux deux derniers mois ne peut pas être imputable à des taux de réponse au-dessus de la moyenne, si on suppose que  $(\gamma_1 + \beta_1)$  est constant.

L'ajustement du poids pour compenser la non-réponse en fonction des chiffres sur les personnes a tendance à augmenter les valeurs de l'indice au premier mois et à les diminuer, du troisième au sixième mois. Les valeurs ainsi calculées pour le premier mois sont généralement plus élevées que celles obtenues à partir de l'ajustement au niveau des ménages. L'ajustement fait avec les chiffres sur les personnes semble corriger les effets que les différences dans les taux de réponse et le nombre de personnes dénombrées dans les groupes de renouvellement ont sur les estimations. Les taux de réponse et le nombre de personnes prises en compte sont faibles au premier mois, ce qui produit une augmentation du biais relatif après l'ajustement. Il semble que la diminution du biais relatif, du troisième au sixième mois, soit due à des taux de réponse

## 5. ANALYSE DES DONNÉES SUR L'INDICE DU BIAIS DE RENOUVELLEMENT

Cette section présente et analyse des tableaux où figurent des données sur l'indice du biais de renouvellement pour la population selon les catégories d'activité sur le marché du travail et en fonction du type de région et des groupes d'âge et de sexe. On obtient les valeurs présentées dans les tableaux en utilisant les poids finals et en effectuant la même correction de la non-réponse à partir des chiffres sur les ménages et des chiffres sur les personnes. On compare les valeurs des indices calculées avec et sans l'ajustement des poids dans le but d'évaluer l'effet de cette correction sur les estimations du biais de renouvellement. L'ajustement du poids des groupes de renouvellement à partir des chiffres sur les ménages a été fait au niveau de chaque province. Les poids finals des ménages dans les six groupes de renouvellement ont donc été multipliés par les facteurs d'ajustement  $R_H(i)$ ;  $i = 1, 2, \dots, 6$ . De même, l'ajustement en fonction des chiffres sur les personnes a été effectué pour chaque province au moyen des facteurs  $R_D(i)$ ;  $i = 1, 2, \dots, 6$ . Pour mesurer l'incidence de ces ajustements sur les estimations calculées pour la population, nous présentons le tableau 3, qui montre l'indice du biais de renouvellement, pour les estimations relatives à la population, selon le type de région et le nombre de mois d'inclusion dans l'échantillon, pour les douze enquêtes menées en 1981. Les valeurs de l'indice calculées sans l'ajustement du poids indiquent un sous-dénombrement relatif des personnes, au premier et au sixième mois, dans les régions AR et NAR. Les valeurs de l'indice obtenues pour les estimations calculées avec l'ajustement des chiffres sur les ménages montrent une amélioration du dénombrement. Toutefois, cet ajustement repose sur l'hypothèse de l'égalité de la taille des ménages d'un groupe de renouvellement à l'autre. Les valeurs de l'indice basées sur l'ajustement par les chiffres sur les répondants se rapprochent plus de 100.0 dans les deux types de région que les valeurs produites par l'ajustement au niveau des ménages. L'ajustement fait à partir des chiffres sur les personnes semble donc mieux corriger les différences dans le dénombrement qui agissent sur les estimations que l'ajustement fondé sur les chiffres concernant les ménages. Le fait que les valeurs de l'indice sont élevées dans les premiers mois et relativement faibles dans les derniers mois pourrait être dû à des variations de la taille des ménages de non-réponse en fonction du nombre de mois d'inclusion dans l'échantillon.



Si  $\hat{Y}(i)$  est un total estimé pour le  $i^{\text{ème}}$  groupe de renouvellement et si  $Y(i)$  représente la vraie valeur du total du  $i^{\text{ème}}$  groupe, l'estimation du biais relatif du total estimé pour le  $i^{\text{ème}}$  groupe de renouvellement peut être calculée par l'équation suivante:

$$B_Y(i) = \frac{\hat{Y}(i) - Y(i)}{Y(i)} ; i = 1, 2, \dots, 6. \quad (6)$$

Étant donné que les  $Y(i)$  sont inconnus et qu'on peut supposer qu'ils sont approximativement égaux (parce que la valeur prévue de la taille des groupes de renouvellement est égale d'un groupe à l'autre dans le cas de grandes régions),  $\frac{\hat{Y}(.)}{Y(.)}$ , la moyenne des estimations d'un total calculées pour les six groupes de renouvellement, peut remplacer  $Y(i)$ .

L'indice du biais de renouvellement pour le  $i^{\text{ème}}$  groupe de renouvellement est défini par l'équation suivante:

$$I_Y(i) = \frac{\hat{Y}(.)}{Y(.)} \cdot 100 = 1 + B_Y(i) \cdot 100 \quad (7)$$

On peut souligner que, puisque la moyenne des estimations d'un total calculées pour les six groupes de renouvellement est utilisée à la place des vraies valeurs,  $I_Y(i)$  peut être biaisé, mais il est utile comme mesure d'évaluation de la différence entre le biais relatif d'un groupe de renouvellement à l'autre pour divers sous-groupes de la population, de même que pour ajuster le poids en fonction des chiffres concernant les ménages et les personnes. De même, il est possible de définir  $P_Y(i)$ , le biais de renouvellement dans les estimations calculées pour la population, pour chacun des groupes de renouvellement.

Lorsque la valeur de l'indice  $I_Y(i)$  est supérieure à 100, le biais relatif est positif, et quand cette valeur est inférieure à 100, le biais relatif est négatif. On peut interpréter l'indice  $I_P(i)$  de la même manière.

est identique, sauf pour quelques exceptions; il y a ainsi vingt régions à l'échelle du Canada qui ont le même facteur de pondération mathématique. Le facteur de sous-pondération de grappe correspond à la fraction de sondage inverse de chaque grappe. Le facteur de compensation corrige le poids pour tenir compte de la non-réponse, et le facteur âge-sexe est une estimation par quotient calculée à partir de projections du nombre total de personnes classées dans les diverses catégories d'âge et de sexe dans chacune des provinces.

Comme il a été expliqué à la section 2, la correction du poids pour tenir compte de la non-réponse est effectuée, pour l'ensemble des ménages dans l'échantillon, à l'intérieur d'unités de compensation. Afin d'évaluer l'effet de l'ajustement du poids par groupe de renouvellement, on a décidé d'utiliser des régions de plus en plus petites (comme unités de compensation), en commençant par les groupes de renouvellement au niveau des provinces. Pour ajuster le poids final dans chacun des groupes de renouvellement dans ces régions, on a multiplié le poids par des facteurs d'ajustement:

$$R_H(i) = \frac{\text{ménages répondants dans l'échantillon}}{\text{ménages répondants dans le groupe de renouvellement (i)}}$$

$$R_P(i) = \frac{\text{personnes répondantes dans l'échantillon}}{\text{personnes répondantes dans le groupe de renouvellement (i)}}$$

Le premier de ces facteurs pondère l'estimation relative aux ménages d'un groupe de renouvellement jusqu'au niveau de l'échantillon des ménages répondants. Le facteur de compensation pondère l'estimation jusqu'au niveau de l'échantillon des ménages dans l'unité de compensation. Le deuxième facteur, calculé à partir du nombre de répondants individuels, pondère l'estimation jusqu'au niveau de l'ensemble de l'échantillon des personnes qui ont répondu à l'enquête, compensant ainsi les différences entre la taille des ménages ou les différences entre le nombre de personnes prises en compte à l'intérieur des ménages. Il est connu que la taille des ménages non répondants a tendance à être plus petite que celle des ménages répondants. Les variations des taux de non-réponse d'un groupe de renouvellement à l'autre peuvent provenir des différences dans la taille moyenne des ménages.



de mois à l'autre, on ne distingue aucune tendance particulière dans l'évolution des proportions de chefs occupés ou en chômage parmi les ménages répondants et non répondants.

L'incidence du premier mois dans la première composante est négative pour tous les mois civils, tant dans le cas des personnes occupées que dans le cas des chômeurs, ce qui indique que le biais du premier mois d'inclusion dans l'échantillon devrait augmenter après l'application de l'ajustement pour compenser la non-réponse.

L'analyse présentée aux sections 2 et 3 désigne les groupes de renouvellement comme ceux auxquels serait appliqué l'ajustement de la non-réponse. Il se peut qu'avec des données réelles on ne puisse pas observer des changements relatifs, à cause de l'effet produit par des différences entre les taux de non-réponse des autres groupes et parce que les valeurs de  $V_{1i}$  et de  $B_i$  peuvent varier pendant la période de six mois. À la section 5, nous analysons l'effet de l'ajustement pour compenser la non-réponse appliqué aux groupes de renouvellement sur le biais de renouvellement, et on tente d'expliquer les résultats à l'aide du modèle décrit plus haut.

Il convient de souligner que l'ajustement de la non-réponse dans le système de pondération actuel de l'EPA se fait au niveau d'unités de compensation qui sont beaucoup plus petites que les UNAR et les UAR délimitées à l'intérieur d'une province. Les estimations du biais de renouvellement qui sont basées sur le système actuel de pondération et de correction de la non-réponse renferment donc un ajustement des différences entre les taux de non-réponse des deux types de région, mais non de celles entre les taux des groupes de renouvellement.

#### 4. CORRECTION DU POIDS AU NIVEAU DES GROUPES DE RENOUVELLEMENT

Le poids final de l'EPA est composé de cinq facteurs: 1) un facteur de pondération mathématique, 2) un facteur rural-urbain, 3) un facteur de sous-pondération de grappe, 4) un facteur de compensation et 5) un facteur âge-sexe. Le facteur de pondération mathématique appliqué à un ménage est égal à l'inverse de la fraction de sondage déterminée par le plan de sondage. Dans chaque province, le poids attribué aux strates urbaines (AR) et rurales (NAR)

peut constater que ces taux varient beaucoup selon le type de région et, pour un type de région donné, les UAR et au niveau du Canada, les taux de non-réponse sont élevés au premier mois, accusent une baisse considérable au deuxième mois et diminuent progressivement au cours des mois suivants. L'ampleur des taux de non-réponse au premier mois est due aux ménages "temporairement absents" et aux cas de "personne à la maison". Par la suite, ces taux diminuent au fur et à mesure que les interviewers déterminent quel est le meilleur moment de communiquer avec ces types de ménages. Les taux de non-réponse sont plus élevés dans les UAR, surtout dans les appartements (qui ne figurent pas au tableau 1), que dans les UNAR. À l'étape du traitement des données, les renseignements fournis par environ 2 % des ménages au cours du mois précédent sont reportés à mois en cours. On obtient les taux de non-réponse présentés dans les tableaux en considérant ces ménages comme des répondants. Le tableau 1 permet également de voir que l'écart entre les taux de non-réponse et leur moyenne,  $(R_1 - \bar{R})$ , est négatif au premier et, parfois, au deuxième mois d'inclusion dans l'échantillon, et positif dans les mois qui suivent. Le taux moyen  $\bar{R}$  est approximativement égal à  $R_2$ . Donc, d'après l'équation (4), le biais relatif au premier mois d'inclusion dans l'échantillon devrait augmenter si  $(\bar{Y}_1 + \bar{\beta}_1)$  et la moyenne de la population  $\bar{Y}_1$  sont supposés invariables. Du troisième au sixième mois, le biais relatif devrait diminuer après la correction du poids pour compenser la non-réponse.

Au tableau 2 figurent les proportions estimées,  $\hat{Y}_1$  et  $\hat{Y}_2$ , des chefs de ménage qui sont occupés ou en chômage, selon le nombre de mois d'inclusion dans l'échantillon, dans le cas des ménages répondants et non répondants. Les estimations ont été calculées à partir des fichiers longitudinaux de l'EPA pour la période de mars à août 1976 et elles sont basées sur des chiffres non pondérés. Les données sur les non-répondants, qui ont participé à l'enquête au moins une fois au cours de la période de six mois, proviennent des mois où ils ont effectivement répondu. Les ménages de non-réponse ont tendance à avoir une plus forte proportion de chefs occupés et une plus faible proportion de chefs en chômage que les ménages répondants. Il est connu que la différence entre les proportions de répondants et de non-répondants classés comme personnes occupées a tendance à être 0.10 et, dans le cas des chômeurs, environ 0.005, le signe de ces différences demeurant identique. D'un nombre

autoreprésentatives (UAR) urbaines et trois ou quatre degrés dans les unités non autoreprésentatives (UNAR) rurales. Aux premières étapes de l'échantillonnage, on applique une méthode de sélection avec probabilité proportionnelle à la taille de la population et, à l'étape finale, où les logements sont sélectionnés parmi les grappes, l'échantillonnage est systématique. A l'intérieur de chaque strate, six numéros de renouvellement sont attribués indépendamment aux grappes sélectionnées. Tous les mois, chaque sixième des ménages font partie de l'enquête depuis 1 à 6 mois. Ainsi, l'échantillon global est composé de six sous-échantillons de même taille qui sont aussi représentatifs les uns que les autres [4]. Il est possible de convertir le numéro de renouvellement des six groupes de renouvellement en un "nombre de mois d'inclusion dans l'échantillon" par une transformation simple.

L'ajustement du poids pour compenser la non-réponse est appliqué à la partie de l'échantillon contenue dans une unité de compensation en fonction du rapport entre le nombre de ménages dans cette unité et le nombre de ménages qui ont répondu à l'enquête. Dans les secteurs NAR, chaque unité primaire d'échantillonnage (UPC) est divisée en deux unités de compensation qui représentent une partie urbaine et une partie rurale. Dans les secteurs AR dupliés d'enquête, les strates (désignées sous-unités) forment des unités de compensation. Le nombre d'unités de compensation dépasse donc 900 dans les UNAR et 800 dans les UAR.

Afin d'évaluer, avec et sans ajustement, le biais dû au renouvellement dans les estimations de l'EPA, on présente et on analyse à la section 3 les données recueillies, au cours des douze mois d'enquête de 1981, sur les taux de non-réponse ( $1-R_1$ ) et sur  $\bar{Y}_1$  et  $\bar{Y}_2$ , les proportions de répondants et de non-répondants classés comme "personnes occupées" et comme "chômeurs". Le "nombre de mois d'inclusion dans l'échantillon" correspond au nombre de mois (y compris le mois en cours) de participation d'un groupe de renouvellement à l'enquête. Aucune donnée sur le biais de réponse,  $\bar{p}_1$ , n'est présentée.

### 3. ANALYSE DES DONNÉES DE L'EPA

Le tableau 1 indique les taux de non-réponse moyens,  $(1-R_1)$ , selon le nombre de mois d'inclusion dans l'échantillon pour les douze mois civils de 1981. On



L'EPA est une enquête mensuelle menée à l'échelle nationale auprès d'un échantillon de 55,000 ménages. Chacune des dix provinces du Canada est divisée en régions économiques, c'est-à-dire en groupes de comtés qui possèdent une structure économique semblable. Ces régions sont décomposées en strates homogènes selon la répartition des personnes occupées dans diverses catégories de professions et d'activités économiques établie d'après les résultats du dernier recensement. On utilise un plan d'échantillonnage stratifié à plusieurs degrés qui prévoit deux degrés d'échantillonnage dans les unités

ment.

Les résultats présentés ci-dessus sont utiles pour évaluer l'importance de divers facteurs dans le biais de renouvellement et pour étudier l'effet de la correction du poids appliquée à ce biais dans chaque groupe de renouvellement.

Les différences entre les taux de réponse de chaque groupe. Les estimations calculées pour un groupe de renouvellement donné s'expliquent par l'ensemble des groupes. Toutefois, le changement introduit dans le biais des produits aucun changement dans le biais des estimations calculées pour l'autre, la correction de la non-réponse dans les groupes de renouvellement ne pour tous les groupes  $i$  et que les taux de réponse varient d'un groupe à l'autre. On peut aussi signaler que, si on suppose que  $\bar{y}_i$  et  $\bar{\beta}_i$  sont constants

Cependant, la différence entre le biais des estimations calculées pour le groupe de renouvellement  $i$  se trouve à l'aide de l'équation (4).

$$B(\bar{y}) - B(\bar{y}_a) = \frac{1}{\sum_{i=1}^K p_i} (R_i - \bar{R}) \bar{\beta}_i. \quad (5)$$

suivante.

différence entre le biais de  $\bar{y}$  et de  $\bar{y}_a$  est le résultat de l'équation biais de réponse,  $\bar{\beta}_i$ , d'un groupe de renouvellement à l'autre. Ainsi, la conditionnement chez les répondants peuvent produire des variations dans le toutefois, les différents degrés de familiarisation avec l'enquête ou de renouvellement ou pour chaque nombre de mois d'inclusion dans l'échantillon. possible que les moyennes  $\bar{y}_i$  soient identiques parmi tous les groupes de

Si les taux de réponse ne varient pas d'un groupe à l'autre, la première composante de l'équation (1) vaut zéro, de sorte que cette équation est identique à (2). Ainsi, un ajustement pour compenser la non-réponse n'entraîne pas une diminution du biais. La différence entre le biais de  $\bar{y}$  et celui de  $\bar{y}_a$  est calculée par l'équation (3).

$$B(\bar{y}) - B(\bar{y}_a) = \frac{1}{K} \sum_{i=1}^K P_i (\bar{R}_i - \bar{R}) (\bar{y}_{1i} + \bar{y}_i) \quad (3)$$

Donc, si les taux de réponse sont différents et si  $\bar{y}_{1i}$  et  $\bar{y}_i$  ne varient pas en fonction du groupe, il n'y a pas de changement dans le biais après l'application de l'ajustement pour compenser la non-réponse à l'intérieur des groupes. Si les moyennes  $\bar{y}_{1i}$  et  $\bar{y}_i$  varient d'un groupe à l'autre, le biais diminue si le membre de droite dans l'équation (3) est positif, mais il augmente si ce membre est négatif. Le changement du biais absolu de  $|B(\bar{y})|$  à  $|B(\bar{y}_a)|$  par suite de la correction de la non-réponse dépend donc du signe et de la grandeur du membre de droite de l'équation (3).

On peut obtenir le biais de l'estimation de la moyenne pour le  $i$ ème groupe de renouvellement, avec et sans ajustement de la pondération pour compenser la non-réponse dans chaque groupe, à l'aide des équations (1) et (2) simplement en fixant  $P_i = 1$  et en conservant les termes qui correspondent aux divers groupes de renouvellement. De plus, l'équation (3) permet de définir par la formule (4) la différence entre le biais des deux estimations calculées pour le  $i$ ème groupe de renouvellement:

$$B(\bar{y}_i) - B(\bar{y}_{ia}) = \left( \bar{R}_i - \bar{R} \right) (\bar{y}_{1i} + \bar{y}_i) \quad (4)$$

où  $\bar{y}_i$  et  $\bar{y}_{ia}$  sont respectivement les estimations obtenues pour le  $i$ ème groupe de renouvellement avant et après l'ajustement. Supposons que  $(\bar{y}_{1i} + \bar{y}_i) > 0$  pour tous les groupes  $i$ ; si  $\bar{R}_i < \bar{R}$ , le biais associé au  $i$ ème groupe de renouvellement augmente après l'ajustement et si  $\bar{R}_i > \bar{R}$ , le biais diminue.

Comme la population de répondants dans divers groupes de renouvellement est identique pendant un mois d'enquête donné, on peut prétendre qu'il est



$i$ ème groupe,  $P_i$  est la proportion de l'ensemble de la population dans le  $i$ ème groupe,  $\bar{p}_i$  est la moyenne du biais de réponse dans le  $i$ ème groupe et  $\bar{R} = \sum_{i=1}^K P_i R_i$ , le taux de réponse global.

L'équation (1) montre comment le biais se décompose en trois composantes. La première composante indique le biais produit par les différences entre les taux de réponse, la deuxième, le biais causé par les différences entre les caractéristiques des répondants et des non-répondants, tandis que la troisième correspond au biais de réponse. Pour simplifier l'analyse, la présente étude utilise des caractéristiques basées sur des attributs comme, par exemple, les proportions de personnes occupées et de chômeurs. L'estimation  $Y_i$ , qui renferme un ajustement pour compenser la non-réponse appliqué à chaque groupe, est exprimée de la façon suivante:

$$\bar{Y}_a = \frac{1}{K} \sum_{i=1}^K \frac{n_i}{n} \bar{Y}_i,$$

où  $n_i$  est la taille de l'échantillon dans le  $i$ ème groupe et  $\bar{Y}_i$  est la moyenne des  $n_i$  unités dans le  $i$ ème groupe. L'équation suivante donne le biais de  $\bar{Y}_a$ :

$$B(\bar{Y}_a) = \sum_{i=1}^K P_i (1-R_i) (\bar{Y}_{1i} - \bar{Y}_{2i}) + \sum_{i=1}^K P_i \bar{p}_i. \quad (2)$$

La première composante de la formule (1), le biais dû aux variations entre les taux de réponse d'un groupe à l'autre, est éliminée, la deuxième composante, le biais produit par la différences entre les caractéristiques des répondants et des non-répondants, reste identique, tandis que la troisième composante, le biais de réponse, peut ne pas avoir la même valeur que dans l'équation (1).

Au moyen d'un modèle qui distingue les erreurs de réponse et de non-réponse et qui utilise les probabilités de réponse au niveau des unités, on a décomposé le biais en une composante attribuable à la non-réponse et une autre causée par la réponse [3]. La décomposition du biais présentée plus haut n'inclut pas les probabilités de réponse des unités individuelles, mais elle est assez simple pour permettre une évaluation empirique des composantes.

## 2. MODÈLE STATISTIQUE

Cette section expose un modèle qui contient des expressions représentant le biais introduit par les différences entre les taux de non-réponse et par les différences entre les caractéristiques des répondants et des non-répondants, ainsi que le biais de réponse de tout groupe dans l'échantillon auquel il est possible d'appliquer un ajustement de la pondération pour compenser la non-réponse. Les groupes de renouvellement peuvent être considérés comme un cas particulier de ce type de groupe.

Supposons qu'une population de taille  $N$  est divisée en "strates" de répondants et de non-répondants de taille  $N_1$  et  $N_2$  respectivement. Un échantillon aléatoire simple de taille  $n$  est prélevé et des réponses sont fournies par  $n_1$  unités, mais non par  $(n - n_1)$  unités.

Supposons aussi que l'échantillon peut être divisé en  $K$  groupes, de sorte que les taux de non-réponse et les caractéristiques des répondants et des non-répondants varient d'un groupe à l'autre. Les méthodes de collecte des données utilisées pour ces groupes et le degré de conditionnement ou de familiarisation des répondants avec l'enquête peuvent être différents, entraînant ainsi des écarts entre les taux de non-réponse, entre les caractéristiques et peut-être aussi dans le biais de réponse. Par l'adaptation d'un résultat présenté dans [2] et [6] de façon à inclure une composante pour le biais de réponse, le biais de la moyenne  $\bar{y}$  de l'échantillon formé de  $n_1$  unités (sans ajustement de la pondération pour compenser la non-réponse dans chaque groupe) correspond à la formule:

$$B(\bar{y}) = \frac{1}{K} \sum_{i=1}^K P_i \bar{y}_i (R_i - \bar{R}) + \frac{1}{K} \sum_{i=1}^K P_i (1 - R_i) (\bar{y}_i - \bar{y}_{2i})$$

où  $\bar{y}_{1i}$  et  $\bar{y}_{2i}$  sont les moyennes de la population de répondants et de non-répondants dans le  $i$ ème groupe,  $R_i$  correspond au taux de réponse du

Le biais de renouvellement peut être dû à plusieurs facteurs. Dans l'EPA, on a observé des variations du taux de non-réponse au niveau des ménages en fonction du groupe de renouvellement, c'est-à-dire selon le nombre de mois de participation des ménages à l'enquête. Il a aussi été souligné que les caractéristiques des ménages de non-réponse ont tendance à être différentes de celles des ménages répondants. Ces deux facteurs peuvent influencer sur le biais. Le facteur de conditionnement du répondant ou sa familiarisation avec l'enquête au cours d'une période de six mois peut faire varier l'importance du biais de réponse dans les données recueillies d'un mois à l'autre. Les données obtenues lors des réentrevues de l'EPA offrent quelques indices de ce genre de biais variable pendant les six mois de participation à l'enquête. Toutefois, selon une hypothèse qui a été formulée dans les études sur le biais, le biais de renouvellement pourrait être causé par les écarts entre les probabilités de non-réponse des groupes de renouvellement [7]. Même si les probabilités individuelles sont inconnues, il est possible d'estimer leur moyenne à partir des taux de non-réponse.

Dans la présente étude, on tente d'évaluer l'effet des différences entre les taux de non-réponse des groupes de renouvellement et entre certaines caractéristiques des répondants et des non-répondants sur le biais de renouvellement. La section 2 présente quelques résultats concernant le biais et décrit l'incidence de ces résultats sur le biais dans les estimations calculées pour différentes groupes de renouvellement. La section 3 fournit quelques données sur les taux de non-réponse dans l'EPA, sur des caractéristiques des répondants et des non-répondants selon le nombre de mois d'inclusion dans l'échantillon et sur l'influence de ces variables sur le biais attribuable au renouvellement. La section 4 explique l'ajustement des poids de l'EPA appliqué aux groupes de renouvellement pour compenser la non-réponse et l'effet de cette correction sur le biais de renouvellement. On y décrit aussi l'indice qui sert de mesure de ce biais. La section 5 contient une analyse des valeurs de cet indice calculées, à partir des enquêtes menées en 1981, pour les diverses catégories d'activité.

## LE BIAIS DE RENOUVELLEMENT DE L'ÉCHANTILLON DANS LES ESTIMATIONS DE L'EPA<sup>1</sup>

P.D. GHANGURDE<sup>2</sup>

L'auteur essaie d'évaluer l'impact d'un ajustement pour la non-réponse, par groupe de renouvellement, sur le biais de renouvellement dans les estimations de l'Enquête sur la population active du Canada. Des résultats sur le biais et sur des caractéristiques des non-répondants sont présentés et discutés. Un indice utilisé pour mesurer le biais de renouvellement est donné et des résultats d'une étude empirique sont analysés.

## 1. INTRODUCTION

Dans le plan d'échantillonnage de l'Enquête sur la population active du Canada (EPA), un sixième des ménages sont supprimés de l'échantillon à chaque mois et la même proportion d'autres ménages viennent les remplacer. L'échantillon est ainsi composé de six panels ou groupes de renouvellement. À un mois donné, les ménages à l'intérieur d'un groupe de renouvellement ont participé à l'enquête depuis un nombre de mois variant d'un à six, y compris le mois en question. Il est bien connu que, dans les enquêtes-ménages fondées sur un plan de renouvellement des unités de l'échantillon, les estimations relatives aux mêmes caractéristiques, qui sont calculées pour différents groupes de renouvellement, peuvent ne pas avoir la même valeur prévue. Ce phénomène, qu'on appelle le biais de renouvellement, a été étudié relativement à l'EPA et à d'autres enquêtes-ménages basées sur un plan de renouvellement des unités de l'échantillon (voir [1], [5], [7] et [8]).

<sup>1</sup> Texte présenté à la réunion de l'American Statistical Association, à Cincinnati, en août 1982.  
<sup>2</sup> P.D. Ghangurde, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.



TABLEAU 9c. Taux de non-réponse des ménages selon le type de région, la taille du ménage et le nombre de mois d'inclusion dans l'échantillon, moyennes annuelles de 1980 et 1981, Canada

Type de région	Nombre de mois					
	Taille du ménage		urbaines		rurales	
	UNAR	UNAR	UNAR	UNAR	UNAR	UNAR
1980	1981					

1	13.99	11.02	11.62	13.57	11.79	9.37
2	7.82	7.31	7.05	7.98	6.84	5.90
3	4.98	6.05	5.03	4.93	5.62	5.02
4	4.03	4.48	4.15	3.59	3.76	3.45
5+	3.13	4.01	3.45	3.21	3.10	2.61
Total	5.32	6.65	5.79	7.25	6.42	4.95
1	8.21	6.79	7.20	6.30	6.32	6.56
2	4.30	4.41	4.13	3.96	3.77	3.47
3	2.44	3.12	2.82	2.24	2.24	2.69
4	1.90	2.48	1.96	1.41	2.56	2.31
5+	1.63	2.23	1.71	1.53	1.97	1.36
Total	3.98	3.93	3.26	3.59	3.50	3.01
1	6.68	6.33	6.50	6.33	5.61	5.52
2	3.83	3.67	3.88	3.44	3.20	2.81
3	2.48	3.18	2.78	1.58	1.32	2.10
4	2.08	2.60	2.27	1.46	1.21	1.93
5+	1.35	1.81	1.87	1.35	2.16	1.37
Total	3.53	3.60	3.20	3.13	2.81	2.52
1	6.41	6.58	6.04	5.50	5.26	5.02
2	4.00	3.63	3.77	3.06	3.11	2.90
3	2.44	2.71	2.91	2.05	1.99	1.87
4	2.05	2.82	2.07	1.74	1.66	1.79
5+	1.47	1.85	1.90	1.51	1.59	1.23
Total	3.53	3.61	3.11	2.99	2.83	2.39
1	6.04	5.94	5.51	4.59	4.47	4.39
2	4.05	3.81	3.81	2.79	3.15	2.55
3	2.51	2.79	2.99	1.60	2.18	1.86
4	2.22	2.56	2.34	1.40	2.10	1.52
5+	1.41	1.75	1.94	1.64	1.17	1.12
Total	3.51	3.50	3.14	2.59	2.75	2.14
1	5.12	5.54	5.30	3.92	3.48	4.24
2	3.43	4.08	3.25	2.42	3.06	2.60
3	2.45	2.51	2.95	1.58	1.48	1.63
4	1.99	1.93	2.07	1.52	1.58	1.86
5+	1.61	1.28	1.68	1.55	1.19	1.21
Total	3.11	3.26	2.85	2.34	2.32	2.19



d'inclusion dans l'échantillon, moyenne annuelle de 1980, Canada.

- 76 -

TABLEAU 9a. Répartition en pourcentage des ménages selon la taille du ménage, le genre de réponse, le type de région et le nombre de mois d'inclusion dans l'échantillon, moyenne annuelle de 1980, Canada.

Nombre de mois	Taille du ménage	UAR		UNAR urbaines		UNAR rurales		Total	
		Ménages répondants	Ménages non répondants	Ménages répondants	Ménages non répondants	Ménages répondants	Ménages non répondants	Ménages répondants	Ménages non répondants
1	1	19.0	39.7	17.8	30.9	10.3	22.1	17.4	36.1
	2	29.4	31.9	29.6	32.8	27.6	34.0	29.1	32.3
	3	18.3	12.3	17.3	14.7	18.7	16.1	18.3	13.2
	4	19.1	10.3	19.0	12.5	21.6	15.2	19.5	11.2
	5+	14.2	5.9	16.3	9.2	21.9	12.7	15.7	7.2
Total		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	1	19.8	42.7	18.1	32.3	10.7	24.7	18.1	38.9
	2	29.5	32.0	29.4	33.2	27.7	35.4	29.2	32.7
	3	18.2	11.0	17.5	13.8	18.6	16.0	18.2	12.1
	4	18.7	8.7	18.9	11.8	21.3	12.7	19.1	9.6
	5+	13.8	5.5	16.1	9.0	21.8	11.2	15.4	6.7
Total		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
3	1	20.1	39.4	18.4	33.4	10.8	22.7	18.4	35.9
	2	29.6	32.1	29.6	30.2	27.7	33.8	29.2	32.2
	3	18.0	12.6	17.1	15.0	18.6	16.1	18.1	13.5
	4	18.7	10.9	19.1	13.7	21.6	14.9	19.1	11.8
	5+	13.6	5.1	15.8	7.8	21.2	12.5	15.2	6.6
Total		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
4	1	20.3	37.9	18.6	35.0	11.0	22.0	18.5	34.9
	2	29.5	33.6	29.4	29.6	27.7	33.7	29.1	33.2
	3	18.1	12.4	17.3	12.8	18.7	17.4	18.2	13.4
	4	18.6	10.6	19.0	14.7	21.2	14.0	19.1	11.7
	5+	13.5	5.5	15.7	7.9	21.4	12.9	15.1	6.9
Total		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
5	1	20.4	36.1	19.1	33.3	11.0	19.8	18.7	32.9
	2	29.5	34.3	29.1	31.7	27.7	33.9	29.1	33.9
	3	18.1	12.8	17.2	13.6	18.6	17.7	18.1	13.8
	4	18.4	11.5	19.1	13.8	21.4	15.7	19.0	12.5
	5+	13.5	5.2	15.6	7.7	21.2	12.9	15.0	6.9
Total		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
6	1	20.6	34.7	19.2	33.3	11.2	21.3	18.9	32.1
	2	29.7	32.8	28.9	36.5	27.6	31.6	29.2	33.0
	3	18.1	14.1	17.1	13.0	18.6	19.3	18.1	14.9
	4	18.4	11.6	19.2	11.2	21.4	15.5	19.0	12.4
	5+	13.2	6.7	15.6	6.0	21.2	12.3	14.9	7.7
Total		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

TABLERAU 8b. Taux de chômage selon la taille du ménage, le nombre de mois d'inclusion dans l'échantillon et le genre de réponse, moyennes annuelles de 1980 et 1981, Canada

Nombre de mois dans l'échantillon		Taille du Genre de ménage					
		1	2	3	4	5	6
1980	1	Répondants Non-répondants	5.88 6.36	5.74 7.16	5.77 10.16	5.83 9.44	5.99 9.29
	2	Répondants Non-répondants	6.94 6.39	6.95 6.81	6.46 7.58	6.54 8.77	7.01 7.97
	3	Répondants Non-répondants	8.14 6.75	7.76 10.37	7.71 7.74	7.71 8.27	8.04 8.79
	4	Répondants Non-répondants	7.54 6.71	7.24 7.42	6.55 10.09	6.84 7.22	6.71 9.05
	5+	Répondants Non-répondants	9.21 8.92	9.10 7.49	8.88 6.50	8.54 9.19	8.57 12.85
	Total	Répondants Non-répondants	7.83 6.82	7.63 7.63	7.27 8.43	7.28 8.60	7.37 9.13
	1981						
	1	Répondants Non-répondants	6.56 5.52	5.86 6.07	6.22 5.29	5.87 9.46	6.16 7.53
	2	Répondants Non-répondants	6.57 5.51	6.38 6.93	6.11 7.18	6.13 7.69	6.37 9.06
	3	Répondants Non-répondants	7.77 6.76	7.59 7.46	7.35 9.19	7.31 8.12	7.57 10.33
1981	4	Répondants Non-répondants	7.30 6.28	7.02 6.05	7.28 5.88	7.02 8.07	6.87 9.77
	5+	Répondants Non-répondants	9.24 7.16	8.90 8.58	8.78 8.27	8.57 8.35	8.26 10.84
	Total	Répondants Non-répondants	7.64 6.05	7.33 6.89	7.28 6.95	7.14 8.30	7.16 9.36
	1	Répondants Non-répondants	6.56 5.52	5.86 6.07	6.22 5.29	5.87 9.46	6.16 7.53
	2	Répondants Non-répondants	6.57 5.51	6.38 6.93	6.11 7.18	6.13 7.69	6.37 9.06
	3	Répondants Non-répondants	7.77 6.76	7.59 7.46	7.35 9.19	7.31 8.12	7.57 10.33
	4	Répondants Non-répondants	7.30 6.28	7.02 6.05	7.28 5.88	7.02 8.07	6.87 9.77
	5+	Répondants Non-répondants	9.24 7.16	8.90 8.58	8.78 8.27	8.57 8.35	8.26 10.84
	Total	Répondants Non-répondants	7.64 6.05	7.33 6.89	7.28 6.95	7.14 8.30	7.16 9.36
	1981						

TABLEAU 8a. Répartition en pourcentage des personnes en fonction de l'activité, selon le nombre de mois d'inclusion dans l'échantillon, la taille du ménage et le genre de réponse, moyennes annuelles de 1980 et 1981, Canada

		Nombre de mois dans l'échantillon																	
		1			2			3			4			5			6		
Taille du ménage	Genre de réponse	0	C	I	0	C	I	0	C	I	0	C	I	0	C	I	0	C	I
		<u>1980</u>																	
1	Répondants	51.0	3.2	45.8	52.3	3.2	44.5	52.2	3.2	44.6	52.2	3.2	44.5	52.4	3.3	44.5	52.2	3.2	44.6
	Non-répondants	65.0	4.4	30.6	62.8	4.9	32.3	60.2	6.8	33.0	60.0	6.3	33.8	60.9	6.2	32.9	57.0	8.1	34.9
2	Répondants	55.5	4.1	40.3	55.7	4.2	40.2	56.1	3.9	40.1	56.0	3.9	40.0	55.6	4.0	40.4	55.5	4.0	40.5
	Non-répondants	58.0	4.0	38.0	56.0	39.8	39.8	56.3	4.6	39.1	55.2	5.3	39.5	58.4	5.1	36.5	60.3	4.9	34.8
3	Répondants	61.7	5.5	32.9	61.9	5.2	32.9	61.9	5.2	33.0	61.9	5.2	32.9	61.6	5.4	33.0	61.9	5.2	33.0
	Non-répondants	62.4	4.5	33.1	57.9	6.7	35.4	59.5	5.0	35.5	58.4	5.3	36.3	61.6	5.9	32.4	58.7	6.3	35.0
4	Répondants	64.1	5.2	30.7	64.2	5.0	30.8	65.0	4.6	30.5	64.8	4.8	30.5	65.1	4.7	30.2	65.1	4.6	30.3
	Non-répondants	64.6	4.6	30.8	64.1	5.1	30.7	57.8	6.5	35.7	63.3	4.9	31.8	62.0	6.2	31.8	61.7	8.1	30.2
5+	Répondants	58.0	5.9	36.1	58.0	5.8	36.2	58.1	5.7	36.3	58.4	5.5	36.2	58.6	5.5	35.0	58.4	5.7	35.9
	Non-répondants	56.8	5.6	37.6	57.9	4.7	37.4	57.1	4.0	39.0	58.4	5.9	35.7	56.6	8.4	35.1	60.1	6.2	33.7
Total	Répondants	58.9	5.0	36.1	59.1	4.9	36.0	59.3	4.7	36.0	59.4	4.7	36.0	59.3	4.7	36.0	59.3	4.7	36.0
	Non-répondants	61.0	4.5	34.6	59.2	4.9	36.0	58.0	5.3	36.7	58.4	5.5	36.1	59.8	6.0	34.2	59.7	6.4	33.9
<u>1981</u>																			
1	Répondants	50.5	3.6	46.0	51.9	3.2	44.8	52.1	3.5	44.5	52.2	3.3	44.6	51.9	3.4	44.7	51.7	3.5	44.8
	Non-répondants	63.7	3.7	32.5	63.4	4.1	32.5	62.3	3.5	34.2	60.8	6.4	32.8	59.0	4.8	36.2	60.8	5.6	33.7
2	Répondants	55.7	3.9	40.4	56.1	3.8	40.1	56.1	3.7	40.3	56.0	3.7	40.3	55.8	3.8	40.4	55.7	3.9	40.4
	Non-répondants	60.4	3.5	36.1	56.9	4.2	38.9	58.4	4.5	37.0	56.4	4.7	38.9	59.2	5.9	35.0	56.5	5.0	38.5
3	Répondants	63.3	5.3	31.4	63.4	5.2	31.4	63.9	5.1	31.1	63.6	5.0	31.4	63.3	5.2	31.5	63.2	5.2	31.6
	Non-répondants	66.6	4.8	28.6	65.5	5.3	29.2	63.8	6.5	29.8	65.9	5.8	28.3	61.5	7.13	1.4	66.1	6.7	27.2
4	Répondants	64.8	5.1	30.1	65.1	4.9	30.0	65.1	5.1	29.8	65.4	4.9	29.6	65.6	4.8	29.6	65.6	5.0	29.4
	Non-répondants	63.4	4.2	32.4	66.4	4.3	29.3	61.6	3.9	34.6	63.3	5.6	31.2	62.3	6.8	30.9	63.2	6.9	30.0
5	Répondants	59.5	6.1	34.4	59.6	5.8	34.6	59.8	5.8	34.5	60.2	5.6	34.2	60.4	5.4	34.2	60.2	5.9	34.0
	Non-répondants	62.3	4.8	32.9	60.1	5.6	34.3	60.2	5.4	34.3	58.5	5.3	36.2	61.6	7.5	31.0	61.8	6.7	31.5
Total	Répondants	59.7	4.9	35.3	60.0	4.8	35.3	60.1	4.7	35.1	60.2	4.6	35.2	60.1	4.6	35.2	60.0	4.8	35.2
	Non-répondants	62.8	4.0	33.2	61.3	4.5	34.2	60.7	4.5	34.8	60.2	5.5	34.4	60.3	6.2	33.5	60.7	5.9	33.4

TABLEAU 7b. Répartition en pourcentage et taux de non-réponse des personnes, par groupe d'âge, selon le nombre de mois d'inclusion dans l'échantillon et le genre de réponse, moyennes annuelles de 1981, Canada

Groupe d'âge	Nombre de mois dans l'échantillon					
	1	2	3	4	5	6
Ménages répondants						
0-14	23.4	23.4	23.5	23.6	23.7	23.8
15-19	9.7	9.6	9.5	9.5	9.4	9.3
20-24	9.4	9.4	9.4	9.4	9.4	9.3
25-44	29.5	29.7	29.7	29.6	29.6	29.6
45-64	19.1	19.1	19.0	19.1	19.1	19.1
65+	8.9	8.9	8.9	8.9	8.9	8.9
Total	100.0	100.0	100.0	100.0	100.0	100.0
Ménages non-répondants						
0-14	18.1	18.7	18.3	19.8	19.8	20.1
15-19	6.9	6.2	6.8	6.6	7.3	7.6
20-24	10.9	10.5	11.1	11.0	11.2	10.7
25-44	33.5	33.2	32.0	32.6	33.1	32.6
45-64	20.2	20.5	20.8	19.2	18.9	19.3
65+	10.4	11.0	11.0	10.8	9.8	9.7
Total	100.0	100.0	100.0	100.0	100.0	100.0
Taux de non-réponse						
0-14	3.96	2.03	1.71	1.85	1.64	1.57
15-19	3.67	1.66	1.57	1.55	1.54	1.50
20-24	5.83	2.80	2.55	2.58	2.34	2.12
25-44	5.67	2.82	2.35	2.42	2.20	2.03
45-64	5.32	2.71	2.38	2.21	1.94	1.86
65+	5.88	3.10	2.68	2.66	2.15	2.01
Total	5.05	2.53	2.18	2.20	1.96	1.85



LEAU 7a. Répartition en pourcentage et taux de non-réponse des personnes, par groupe d'âge, selon le nombre de mois d'inclusion dans l'échantillon et le genre de réponse, moyennes annuelles de 1980, Canada

Nombre de mois dans l'échantillon

Groupe d'âge	Nombre de mois dans l'échantillon						Taux de non-réponse					
	1	2	3	4	5	6	Total	-14	-19	-24	-44	+ Total
14-19	24.2	24.1	24.3	24.4	24.5	24.5	100.0	18.9	6.3	7.6	31.9	10.6
19-24	9.9	9.8	9.7	9.7	9.7	9.6	100.0	19.0	19.6	20.4	30.8	10.8
24-44	9.2	9.2	9.2	9.1	9.1	9.0	100.0	10.3	7.8	21.8	31.5	20.6
44-64	18.9	18.9	18.9	18.9	18.9	19.0	100.0	10.1	10.5	21.5	31.4	10.4
64+	29.1	29.1	29.2	29.3	29.2	29.1	100.0	19.5	19.9	20.9	31.5	10.2
Total	8.7	8.8	8.7	8.7	8.7	8.7	100.0	19.4	7.5	10.4	31.6	10.0
14-19	24.2	24.1	24.3	24.4	24.5	24.5	100.0	18.9	6.3	7.6	31.9	10.6
19-24	9.9	9.8	9.7	9.7	9.7	9.6	100.0	19.0	19.6	20.4	30.8	10.8
24-44	9.2	9.2	9.2	9.1	9.1	9.0	100.0	10.3	7.8	21.8	31.5	10.4
44-64	18.9	18.9	18.9	18.9	18.9	19.0	100.0	10.1	10.5	21.5	31.4	10.4
64+	29.1	29.1	29.2	29.3	29.2	29.1	100.0	19.5	19.9	20.9	31.5	10.2
Total	8.7	8.8	8.7	8.7	8.7	8.7	100.0	19.4	7.5	10.4	31.6	10.0

TABLEAU 6b. Répartition en pourcentage et taux de non-réponse des personnes, par groupe d'âge, selon la taille du ménage et le genre de réponse, moyennes annuelles de 1981, Canada

Groupe d'âge	Taille du ménage					Répondants	Non-répondants	Taux de non-réponse	
	1	2	3	4	5				
0-14	0.0	2.7	19.8	34.4	36.9	23.6	0.1	3.8	18.9
15-19	1.8	3.4	8.6	9.9	16.1	9.5	2.0	3.1	6.9
20-24	10.6	14.1	11.4	6.4	7.1	10.9	13.0	15.1	10.9
25-44	28.8	25.9	31.8	35.2	25.8	33.0	39.6	28.6	33.0
45-64	24.4	31.7	23.5	12.6	11.6	19.9	22.9	30.1	19.9
65+	34.3	22.2	5.0	1.5	2.5	10.5	22.3	19.3	10.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0-14	0.1	3.8	23.6	37.2	39.4	18.9	0.1	3.8	18.9
15-19	2.0	3.1	7.5	8.8	15.6	6.9	2.0	3.1	6.9
20-24	13.0	15.1	12.0	5.9	5.8	10.9	13.0	15.1	10.9
25-44	39.6	28.6	34.3	37.2	27.7	33.0	39.6	28.6	33.0
45-64	22.9	30.1	19.8	10.1	10.2	19.9	22.9	30.1	19.9
65+	22.3	19.3	2.8	0.9	1.2	10.5	22.3	19.3	10.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0-14	-	5.16	2.77	2.05	1.77	2.12	-	5.16	2.12
15-19	7.15	3.47	2.03	1.69	1.61	1.92	7.15	3.47	1.92
20-24	7.94	4.02	2.47	1.77	1.37	3.04	7.94	4.02	3.04
25-44	8.85	4.14	2.51	2.01	1.78	2.91	8.85	4.14	2.91
45-64	6.20	3.57	1.97	1.53	1.46	2.74	6.20	3.57	2.74
65+	4.38	3.27	1.31	1.15	0.83	3.08	4.38	3.27	3.08
Total	6.58	3.76	2.33	1.90	1.66	2.63	6.58	3.76	2.63

TABLEAU 6a. Répartition en pourcentage et taux de non-réponse des personnes, par groupe d'âge, selon la taille du ménage et le genre de réponse, moyennes annuelles de 1980, Canada

Groupe d'âge	Taille du ménage					
	1	2	3	4	5+	Total
<hr/>						
Répondants						
0-14	0.0	2.6	20.5	35.1	37.1	24.3
15-19	1.9	3.5	8.2	9.7	16.7	9.7
20-24	10.5	14.2	11.4	6.0	6.7	9.1
25-44	27.9	25.3	31.3	35.4	25.2	29.2
45-64	25.4	31.6	23.5	12.4	11.8	18.9
65+	34.3	22.8	5.2	1.3	2.6	8.7
Total	100.0	100.0	100.0	100.0	100.0	100.0
<hr/>						
Non-répondants						
0-14	0.3	2.6	22.6	37.0	40.8	19.4
15-19	3.0	3.8	8.1	8.8	15.7	7.5
20-24	13.6	14.4	12.0	4.8	5.7	10.4
25-44	35.5	27.3	33.5	37.4	26.6	31.6
45-64	25.2	33.0	19.6	10.7	10.2	20.9
65+	22.4	19.1	4.1	1.4	1.1	10.2
Total	100.0	100.0	100.0	100.0	100.0	100.0
<hr/>						
Taux de non-réponse						
0-14	-	4.52	3.30	2.57	2.05	2.50
15-19	11.07	4.83	2.93	2.22	1.76	2.42
20-24	9.49	7.94	3.16	1.95	1.60	3.54
25-44	9.33	4.83	3.21	2.58	1.98	3.38
45-64	7.45	4.68	2.52	2.11	1.62	3.45
65+	5.01	3.80	2.41	2.47	0.85	3.65
Total	7.48	4.50	3.00	2.44	1.87	3.13

TABLEAU 5. Répartitions en pourcentage des personnes par groupe d'âge et genre de réponse, moyennes annuelles de 1980 et 1981, Canada

Groupe d'âge	1980		1981			
	Répondants	Non-Répondants	Taux de non-réponse	Répondants	Non-répondants	Taux de non-réponse
0-14	24.3	19.4	2.50	23.6	18.9	2.12
15-19	9.7	7.5	2.42	9.5	6.9	1.92
20-24	9.1	10.4	3.54	9.4	10.9	3.04
25-44	29.2	31.6	3.38	29.6	33.0	2.91
45-64	18.9	20.9	3.45	19.1	19.9	2.74
65+	8.7	10.2	3.65	8.9	10.5	3.08
Total	100.0	100.0	3.13	100.0	100.0	2.63

TABLEAU 4a. Définition de la variable type de famille

Code	Nombre de familles économiques dans le ménage	Taille des familles économiques	Âge du chef de famille	Présence d'enfants dans le ménage	Le chef est membre d'un couple marié
------	---	---------------------------------	------------------------	-----------------------------------	--------------------------------------

1	1	1	25		
2	1	1	25-64		
3	1	1	65+		
4	1	1	45	Non	Oui
5	1	1	45	Non	Oui
6	1	2	45	Oui	Oui
7	1	2+	45	Oui	Oui
8	1	2+		Non	Oui
9	1	2+		Non	Non
10	1	2+		Oui	Non
11	2+	toutes, 1 personne			
12	2+	toutes, 2 personnes			
13	2+	et plus familles variées			

TABLEAU 4b. Répartition en pourcentage des ménages répondants et non répondants selon le type de famille économique, moyennes annuelles de 1980 et 1981, Canada

	1980	1981
Type de famille	Ménages non-économique	Ménages non-économique
Taux de	Taux de	Taux de
Ménages	Ménages	Ménages
non-économique	non-économique	non-économique
répondants	répondants	répondants
non-réponse	non-réponse	non-réponse

1	5.8	2.3	9.80	5.7	2.4	7.82
2	21.0	9.5	8.51	23.4	9.9	7.71
3	8.6	6.6	5.14	9.1	7.0	4.47
4	9.5	8.0	4.72	9.5	8.0	4.05
5	15.5	13.6	4.57	14.4	13.8	3.57
6	18.4	28.1	2.67	16.3	27.0	2.10
7	4.9	9.9	2.04	4.7	9.0	1.83
8	4.6	8.2	2.32	4.1	8.6	1.65
9	3.1	4.3	2.99	3.1	4.3	2.45
10	3.9	4.8	3.30	5.0	4.9	3.47
11	3.4	2.7	5.05	3.5	2.8	4.25
12	0.0	0.1	1.25	0.1	0.1	2.29
13	1.2	2.1	2.47	1.2	2.1	2.06
Total	100.0	100.0	100.0	100.0	100.0	



TABLEAU 3. Répartition en pourcentage des ménages répondants et non répondants selon la taille du ménage et le nombre de mois d'inclusion dans l'échantillon, moyennes annuelles de 1980, et 1981, Canada

Nombre de mois	Taille du ménage											
	1980						1981					
	1	2	3	4	5+	Total	1	2	3	4	5+	Total
Ménages répondants												
1	17.4	29.1	18.3	19.5	15.7	100.0	18.3	29.6	17.9	19.5	14.7	100.0
2	18.1	29.2	18.2	19.1	15.4	100.0	19.0	29.8	17.7	19.2	14.3	100.0
3	18.4	29.2	18.1	19.1	15.2	100.0	19.2	29.7	17.8	19.0	14.3	100.0
4	18.5	29.1	18.2	19.1	15.1	100.0	19.5	29.7	17.8	18.9	14.2	100.0
5	18.7	29.1	18.1	19.0	15.0	100.0	19.7	29.7	17.7	18.9	14.0	100.0
6	18.9	29.2	18.1	19.0	14.9	100.0	19.8	29.8	17.7	18.8	13.9	100.0
Ménages non-répondants												
1	36.1	32.3	13.2	11.2	7.2	100.0	37.7	32.9	12.8	10.3	6.3	100.0
2	38.9	32.7	12.1	9.6	6.7	100.0	40.5	32.7	11.5	9.1	6.2	100.0
3	35.9	32.2	13.5	11.8	6.6	100.0	41.5	33.0	9.6	9.4	6.5	100.0
4	34.9	33.2	13.4	11.7	6.9	100.0	37.6	31.7	12.5	11.0	7.1	100.0
5	32.9	33.9	13.8	12.5	6.9	100.0	36.5	32.6	12.0	10.9	8.0	100.0
6	32.1	33.0	14.9	12.4	7.7	100.0	34.1	32.4	12.3	12.7	8.5	100.0
Taux de non-réponse												
1	13.39	7.64	5.10	4.10	3.30	6.94	12.81	7.34	4.85	3.63	2.97	6.66
2	7.90	4.28	2.59	1.97	1.71	3.84	7.03	3.74	2.25	1.65	1.51	3.42
3	6.56	3.81	2.61	2.17	1.54	3.47	6.19	3.28	1.62	1.49	1.37	2.96
4	6.32	3.92	2.56	2.15	1.61	3.45	5.32	3.02	2.01	1.67	1.44	2.83
5	5.87	3.97	2.63	2.28	1.61	3.42	4.50	2.72	1.70	1.45	1.43	2.48
6	5.05	3.41	2.51	2.00	1.59	3.03	3.82	2.45	1.58	1.53	1.39	2.25
Total	7.48	4.50	3.00	2.44	1.89	4.02	6.58	3.76	2.33	1.90	1.69	3.43

TABLEAU 2. Répartitions en pourcentage et taux de non-réponse des ménages répondants et non répondants selon la taille du ménage, moyennes annuelles de 1980 et 1981, Canada

Taille du ménage	1980				1981			
	Total	Répondants	Non-répondants	Taux de non-réponse	Total	Répondants	Non-répondants	Taux de non-réponse
1	19.0	18.3	35.4	7.48	19.9	19.2	38.1	6.58
2	29.3	29.2	32.8	4.50	29.8	29.7	32.6	3.76
3	17.8	18.2	13.4	3.00	17.6	17.8	11.9	2.33
4	18.8	19.1	11.4	2.44	18.8	19.1	10.4	1.90
5+	14.9	15.2	7.0	1.89	14.0	14.2	6.9	1.69
Total	100.0	100.0	100.0	4.02	100.0	100.0	100.0	3.43
Taille moyenne des ménages	2.91	2.93	2.26		2.86	2.88	2.19	

TABLEAU 1. Répartitions en pourcentage des ménages répondants et non-répondants selon le nombre de mois d'inclusion dans l'échantillon, 1980 et 1981, Canada

Nombre de mois	Total	Répondants	Non-répondants	Taux de non-réponse
1980				
1	16.6	16.1	28.6	6.94
2	16.6	16.7	15.9	3.84
3	16.7	16.8	14.4	3.47
4	16.7	16.8	14.3	3.45
5	16.7	16.8	14.2	3.42
6	16.8	16.9	12.6	3.03
Total	100.0	100.0	100.0	4.02
1981				
1	16.6	16.0	32.1	6.66
2	16.7	16.7	16.6	3.42
3	16.7	16.8	14.4	2.96
4	16.7	16.8	13.9	2.83
5	16.7	16.9	12.1	2.48
6	16.7	16.9	11.0	2.25
Total	100.0	100.0	100.0	

constances exceptionnelles à l'intérieur d'un ménage. Les logements vacants incluent les logements inoccupés, les logements saisonniers, les logements en construction, les logements occupés par une personne qui ne doit pas être interviewée, ainsi que les logements démolis, convertis en établissements commerciaux, déplacés, désaffectés (inhabitables) ou ceux qui figurent par erreur sur la liste.

4. Pour des renseignements supplémentaires sur l'ajustement de la non-réponse dans l'EPA, voir Méthodologie de l'enquête sur la population active du Canada (1976), Statistique Canada, n° 71-526 au catalogue, hors série, octobre 1977, p. 67.

5. Pour plus de détails concernant le processus de pondération dans l'EPA, voir Méthodologie de l'enquête sur la population active du Canada (1976), Statistique Canada, n° 71-526 au catalogue, hors série, octobre 1977, p. 65-74.

Il existe un bon nombre d'autres procédés qui pourraient remplacer la méthode actuelle de compensation de la non-réponse. L'élaboration de nouvelles techniques de compensation doit se faire en deux étapes. La première a trait à la simulation et à l'évaluation des estimations mensuelles de la population active calculées par la méthode d'imputation proposée dans la présente étude. La deuxième étape consiste à mettre au point d'autres procédés d'ajustement de la non-réponse, pour ensuite en faire une évaluation empirique. Une telle étude est actuellement en train.

## NOTES DES RENVOIS

1. Les estimations tirées de l'enquête sur la population active concernent la semaine précise visée par l'enquête à chaque mois, la semaine de référence, normalement celle qui comprend le 15<sup>e</sup> jour du mois. La semaine d'enquête, pendant laquelle toutes les interviews ont lieu, est la semaine qui suit immédiatement la semaine de référence.

2. L'univers de l'enquête sur la population active comprend toutes les personnes âgées de 15 ans et plus et qui demeurent au Canada, à l'exception des personnes suivantes: les résidents du Yukon et des Territoires du Nord-Ouest, les personnes vivant dans les réserves indiennes, les pensionnaires d'institutions et les membres à temps plein des Forces armées canadiennes.

3. Tous les mois, l'interviewer doit indiquer a) si une interview complète a eu lieu ou, autrement dit, si un questionnaire de l'enquête sur la population active a été rempli au complet pour tous les membres d'un ménage admissible à l'enquête; b) si une interview partielle a été faite, c'est-à-dire si un questionnaire a été rempli pour quelques-uns des membres admissibles d'un ménage; ou c) si aucune interview n'a eu lieu. Dans ce dernier cas, l'interviewer doit inscrire la raison pour laquelle il n'a pas pu mener d'interview. Les ménages non-répondants comprennent ceux où personne n'était présent au domicile (après plusieurs tentatives de communiquer avec ces ménages), ceux qui refusent de répondre à l'enquête, ceux qui étaient temporairement absents de leur domicile, ainsi que les cas où l'interview n'a pas pu avoir lieu à cause de conditions climatiques défavorables, de décès, de maladie, de problèmes linguistiques ou d'autres circonstances.



ménages répondants. La taille du ménage est donc une variable importante à ajouter à toute technique de compensation de la non-réponse. La présente analyse révèle que les écarts les plus grands entre les répondants et les non-répondants sont observés parmi les ménages qui comptent un membre. Il pourrait donc être possible de faire des corrections pour deux groupes de ménages seulement, à savoir les ménages composés d'un membre et ceux qui comptent deux membres ou plus. L'inclusion de la taille du ménage dans les méthodes de compensation de la non-réponse au niveau des ménages exige que certains renseignements sur la taille des ménages non-répondants soient connus. Ces données peuvent être explicites comme, par exemple, la taille du ménage enregistrée à un mois d'enquête antérieure, ou implicites comme, par exemple, la répartition des ménages non-répondants selon la taille dans les mois d'enquête antérieurs ou la répartition des ménages selon la taille à partir d'une source indépendante comme le recensement. Quoi qu'il en soit, la combinaison d'ajustements prenant en compte la taille du ménage et de corrections effectuées selon le numéro de renouvellement devrait aider à réduire les différences qui existent entre les groupes de renouvellement dans les échantillons utilisés pour produire les estimations relatives aux ménages et aux familles économiques.

Tel qu'il a été expliqué à la section 3.9, même pour une taille de ménage donnée et pour un nombre donné de mois d'inclusion dans l'échantillon, il y a des différences entre les répartitions des répondants et des non-répondants en fonction de l'activité. Pour l'EPA, il pourrait être utile d'intégrer des variables relatives au marché du travail dans le processus d'ajustement. En pratique, cependant, deux facteurs ont tendance à exclure cette possibilité. D'abord, il y a un besoin d'un ajustement global des poids, non seulement dans l'EPA mais aussi dans les diverses enquêtes supplémentaires. Deuxièmement, les renseignements obtenus pour ce degré d'agrégation seraient peu stables et exigeraient des ajustements à des degrés d'agrégation plus élevés. Ce nouveau type d'ajustement ferait disparaître les avantages que pourrait produire l'évolution d'un marché du travail local. Toute technique de compensation doit tenir compte du fait que le taux de non-réponse dans l'EPA est actuellement assez bas, ce qui a un effet sur le degré de complexité requis, sur le calcul des estimations et sur la fiabilité des données utilisées pour compenser la non-réponse, données qui sont un élément clé de la technique appliquée.

par l'inverse du taux de réponse des ménages. Ces corrections sont effectuées à partir des totaux des ménages sans prendre en considération les caractéristiques des ménages. À moins qu'il n'existe une forte corrélation entre les ménages dans une unité de compensation, la technique d'ajustement actuelle ne devrait pas produire une grande diminution du biais dû à la non-réponse.

Il serait souhaitable d'avoir une idée de l'importance du biais de non-réponse qui s'introduit dans la méthode actuelle de compensation de la non-réponse. Il est possible d'imputer les renseignements qui manquent dans le fichier de l'EPA à cause de la non-réponse par des techniques semblables à celles appliquées dans la présente étude. Après des ajustements pour tenir compte de la non-réponse totale (c'est-à-dire la non-réponse pendant les six mois d'inclusion dans l'échantillon), on peut obtenir des estimations d'enquête à l'aide de ces méthodes complètes d'imputation. Une comparaison des estimations ainsi calculées avec les estimations officielles de l'enquête peut fournir des renseignements supplémentaires pour appuyer les conclusions présentées dans cette étude concernant le biais dû à la non-réponse.

La présente étude avance des arguments en faveur de l'intégration de diverses variables additionnelles à la correction de la non-réponse, notamment le nombre de mois d'inclusion dans l'échantillon, la taille du ménage et l'activité sur le marché du travail. Comme il y a des variations considérables du taux de réponse selon le numéro de renouvellement (le nombre de mois d'inclusion dans l'échantillon), il conviendrait d'effectuer les ajustements de la non-réponse à l'intérieur de chaque groupe de renouvellement séparément. Étant donné que les caractéristiques des non-répondants relatives au marché du travail varient dans une certaine mesure d'un mois à l'autre, un ajustement qui tient compte du groupe de renouvellement devrait également améliorer les estimations de la population active. Comme on observe les différences les plus marquées entre le premier mois d'inclusion dans l'échantillon et les mois suivants, un ajustement appliqué à ces deux catégories pourrait être suffisant.

Parmi les ménages non-répondants, on trouve beaucoup plus de ménages formés d'un seul membre (et, dans une moindre mesure, de deux membres) que parmi les

La technique de compensation des ménages non-répondants adoptée dans l'EPA est appliquée à l'intérieur de petites régions géographiques (unités de compensation) par la multiplication du poids calculé d'après le plan d'échantillonnage

utilisées dans l'enquête.

étudiées à la section 3 dans les méthodes de compensation de la non-réponse techniques semblables, il semble important d'intégrer quelques-unes des variables hypothèse que les ménages répondants et non-répondants ont des caractéristiques personnes varient quelque peu selon le genre de réponse. Si on pose comme à toute une gamme de variables. Les caractéristiques des ménages et/ou des La présente étude définit les ménages répondants et non-répondants par rapport

## 5. RÉSUMÉ

Le genre de réponse.

ments supplémentaires utiles à la caractérisation des unités d'enquête selon l'échantillon, la variable type de région n'offre pas beaucoup de renseignement des ménages par la taille et le nombre de mois d'inclusion dans régions qui enregistrent des taux de non-réponse différents. Mais dans un ques de compensation de la non-réponse, puisqu'elle sert à distinguer les régions. La variable type de région est un facteur important dans les techniques de compensation de la non-réponse, puisqu'elle sert à distinguer les mêmes que ceux observés dans l'analyse de ces ménages pour l'ensemble des résultats relatifs aux répondants et aux non-répondants sont généralement chaque type de région. Pour des ménages d'une taille donnée, les écarts entre rences sont une fonction des répartitions des ménages selon la taille et pour dans les UNAR urbaines et sont les plus bas dans les UNAR rurales. Ces différences (taille) sont les plus élevées dans les UR, atteignent un niveau moins élevé Les taux de non-réponse de l'ensemble des ménages (c'est-à-dire peu importe la

subséquents.

Le deuxième mois, et des diminutions un peu moins marquées, au cours des mois fois, on constate une baisse importante de la non-réponse entre le premier et unités (c'est-à-dire les résultats présentés à la section 3.3). Encore une suivant ainsi le même comportement que celui observé pour l'ensemble des diminuent lorsque le nombre de mois d'inclusion dans l'échantillon augmente, Les taux de non-réponse, même s'ils varient d'un type de région à l'autre,



#### 4.9 Type de région

généralement plus élevé que celui des répondants. Les ménages formés d'un seul membre, le taux de chômage des non-répondants est

ménages non-répondants lorsqu'on examine les diverses tailles de ménage. Dans

tendance précise dans l'évolution des taux de chômage dans le temps chez les

lorsqu'on ne tient pas compte de la taille du ménage). On ne distingue aucune

Les résultats décrits à la section 3.3 montrent qu'il existe des différences

notables entre les répartitions des ménages des répondants et non-répondants

établies selon la taille et le nombre de mois d'inclusion dans l'échantillon.

La présente section poursuit l'analyse de ces résultats en fonction de grandes

catégories de régions délimitées généralement selon la concentration et la

densité de la population; ce sont les unités autorenseignantes (UAR), les

unités non autorenseignantes urbaines (UNAR urbaines) et les unités non

autorenseignantes rurales (UNAR rurales). Bien qu'il existe des définitions

plus précises de ces types de région, il suffit de noter que, pour les besoins

de la présente étude, les UAR correspondent aux grandes villes du Canada, les

UNAR urbaines aux petites villes et les UNAR rurales aux régions du pays les

moins peuplées, y compris les petits villages et les terres agricoles. On ne

tient pas compte des régions spéciales en raison de la très petite taille de

l'échantillon. Sur un plan très général, les rapports observés à la

section 3.3 pour l'ensemble des régions correspondent à ceux établis pour les

trois grandes catégories de région. Cependant, la répartition des répondants

en fonction de la taille du ménage varie suivant le type de région. D'après

les répartitions, la plupart des ménages des UAR sont de taille assez petite,

alors que les UNAR rurales comptent beaucoup moins de ménages de petite

taille. Le contraire s'applique aux ménages de grande taille. Le rapport

entre les ménages répondants et non-répondants est toutefois assez semblable

d'un type de région à l'autre. On peut constater aux tableaux 9a et 9b que le

pourcentage de ménages non-répondants répartis dans les ménages formés d'un

seul membre est environ le double du pourcentage de ménages répondants dans la

même catégorie de taille, tandis que la proportion de ménages non-répondants

composés de 5 personnes ou plus représente près de la moitié de celle des mé-

nages répondants ayant cette taille.

Dans le cas des ménages composés de 2, 3, 4 ou 5 personnes ou plus, on constate la même relation générale entre les taux de chômage des ménages répondants et non-répondants que pour l'ensemble des individus (c'est-à-dire

suivants.

premier mois, mais suit une évolution un peu variable au cours des mois chez les répondants dont le taux de chômage est à son plus haut niveau au mois d'inclusion dans l'échantillon augmenté. On n'observe pas ce phénomène a une progression considérable du taux de chômage à mesure que le nombre de système mois varie d'une année à l'autre. Chez les ménages non-répondants, il y quatrième au sixième mois. Le rapport entre les taux du deuxième et du troisième mois des répondants, au premier mois, et supérieur à celui des répondants du quatrième (le taux de chômage des non-répondants est inférieur à la taille du ménage), le rapport entre les taux du deuxième et du troisième mois des personnes soumises à l'enquête (c'est-à-dire peu importe

meurs parmi les répondants et les non-répondants.

des liens qui existent entre les proportions de personnes occupées et de chômage entre les taux de chômage des répondants et des non-répondants découle rapport entre les taux de chômage des répondants et des non-répondants, puisque le tableaux précédents et on peut faire des observations semblables, puis que pour 1980 et 1981 respectivement. Ces résultats s'apparentent à ceux des nombre de mois d'inclusion dans l'échantillon et le genre de réponse, le tableau 8b présente le taux de chômage ventilé par la taille du ménage, le

répondants est l'inverse de celui établi pour les inactifs.

actives, le rapport général entre les répartitions des répondants et des non-personnes occupées représentent la plus grande partie du groupe des personnes 3 personnes ou plus, il n'existe aucune tendance précise. Étant donné que les répondants que parmi les répondants mais, chez les ménages formés de ou deux personnes, il y a moins de personnes inactives dans les ménages non-dants et d'une taille de ménage à l'autre. Parmi les ménages composés d'une Le pourcentage des "inactifs" varie entre les ménages répondants et non-répon-

Le tendance est variable.

qui en sont à leur deuxième ou troisième mois d'inclusion dans l'échantillon, quatrième au sixième mois, que dans les ménages répondants. Quant aux ménages



Étant donné que le pourcentage de chômeurs est plus sensible aux fluctuations de la taille de l'échantillon que les autres formes d'activité sur le marché du travail, et qu'il affiche une tendance nette dans le temps, les corrections de la non-réponse appliquées à l'ensemble des groupes de renouvellement déformeraient cette caractéristique. Un tel ajustement produirait une surestimation du chômage au premier mois et une sous-estimation, du quatrième au sixième mois. Comme l'écart entre la répartition en pourcentage des répondants et des non-répondants dans la catégorie des chômeurs devient plus prononcé au cours des derniers mois d'inclusion dans l'échantillon, l'effet global serait une sous-estimation du chômage. Étant donné que la correction de la non-réponse se fait au niveau des ménages, non au niveau des personnes, et que la taille du ménage s'est révélée un important déterminant de la réponse et de la non-réponse (voir section 3.2), il est essentiel de considérer la taille du ménage comme un élément additionnel pour l'évaluation du rapport entre la non-réponse et l'activité sur le marché du travail.

Lorsqu'on examine les répartitions des personnes en fonction de l'activité, du nombre de mois de participation à l'enquête et de la taille du ménage, les tendances ou les rapports décrits ci-dessus ne s'appliquent pas. Chez les ménages comptant un membre, les proportions de personnes occupées et de chômeurs sont beaucoup plus élevées parmi les non-répondants que parmi les répondants. Chez ces derniers, la proportion de personnes occupées et la proportion de chômeurs sont relativement constantes pour divers nombres de mois d'inclusion dans l'échantillon. Dans le cas des ménages non-répondants, on observe une diminution générale de la proportion de personnes occupées à mesure que le nombre de mois d'inclusion augmente, tandis que la proportion de chômeurs s'accroît considérablement.

En ce qui concerne les ménages qui comptent plus d'un membre (2, 3, 4 et 5 membres ou plus), les différences entre les répartitions des répondants et des non-répondants en fonction de l'activité sont beaucoup plus petites. En outre, le rapport entre ces répartitions est loin d'être aussi fort ou uniforme que dans le cas des ménages formés d'une personne. D'après les répartitions, il y a généralement moins de chômeurs dans les ménages non-répondants, au premier mois d'inclusion dans l'échantillon, et plus de chômeurs du

au tableau 8a la répartition des personnes, selon l'activité, pour chaque catégorie définie par la taille du ménage, le nombre de mois dans l'échantillon et le genre de réponse.

Un examen de la répartition des personnes selon l'activité, sans tenir compte de la taille des ménages révèle des différences marquées entre les ménages répondants et les ménages de non-répondants, et ces différences ne sont pas uniformes dans le temps. C'est peut-être dans les pourcentages des chômeurs qu'on observe les variations les plus intéressantes. Parmi les répondants, ce pourcentage est relativement constant pour chaque nombre de mois d'inclusion dans l'échantillon mais, dans le cas des ménages non-répondants, il y a un accroissement du pourcentage de chômeurs quand le nombre de mois de participation de ces ménages à l'enquête grandit. Le pourcentage de chômeurs (âgés de 15 ans et plus) dans les ménages répondants varie d'un minimum de 4.7 %, du troisième au sixième mois, à un maximum de 5.0 % au premier mois de 1980 et, en 1981, d'un minimum de 4.6 % aux quatrième et cinquième mois à un maximum de 4.9 % au premier mois. Dans le cas des ménages non-répondants, les pourcentages respectifs sont de 4.5 % au premier mois et de 6.4 % au sixième mois de 1980, et de 4.0 % au premier mois et de 6.2 % au cinquième mois de 1981. Une comparaison du pourcentage de chômeurs dans les ménages répondants et non-répondants à différents moments révèle qu'il y a moins de chômeurs parmi les non-répondants que parmi les répondants au premier mois, tandis qu'il y a plus de chômeurs chez les non-répondants que chez les répondants, du quatrième au sixième mois d'inclusion des ménages dans l'échantillon. Mais ce rapport est variable pendant le deuxième et le troisième mois. Une comparaison dans le temps des répartitions en pourcentage des personnes en fonction de l'activité sur le marché du travail révèle que les ménages répondants ont des répartitions assez stationnaires, mais que celles des ménages non-répondants varient. Chez les ménages de non-réponse, on observe plus de fluctuations d'un mois à l'autre dans les répartitions en pourcentage à l'intérieur de chaque activité que chez les répondants. On ne distingue pas de tendance nette dans les changements, sauf pour le chômage, qui grandit avec le nombre de mois d'inclusion dans l'échantillon. Cette variation parmi les non-répondants est attribuable au moins en partie à la petite taille des échantillons de non-répondants comparativement à celle des échantillons de répondants.

tableaux 7a et 7b, pour 1980 et 1981 respectivement, de même que les taux de non-réponse correspondants.

Ces tableaux permettent de constater que les répartitions des répondants par groupes d'âge sont pratiquement identiques, quel que soit le nombre de mois d'inclusion dans l'échantillon. Bien que les répartitions des non-répondants affichent un plus haut degré de variabilité d'un mois à l'autre, ces répartitions s'avèrent quand même assez stables. Les différences entre les répartitions des répondants et des non-répondants pour chaque mois pris séparément correspondent aux totaux obtenus pour l'ensemble des six mois.

Un examen des taux de non-réponse individuels indique encore une fois que ces taux ont généralement tendance à diminuer quand le nombre de mois augmente. Ce comportement est observé tant pour chaque groupe d'âge que pour l'ensemble des enquêtes. Comme prévu, l'évolution dans le temps du taux de non-réponse n'est pas aussi prononcée au niveau des personnes qu'au niveau des ménages. Cela peut être attribué à des changements de comportement des ménages de diverses tailles vis-à-vis de la participation à l'enquête. C'est-à-dire que les ménages de grande taille ont tendance à devenir des non-répondants dans les derniers mois de leur inclusion dans l'enquête, tandis que les ménages de petite taille ont tendance à devenir des répondants (voir le tableau 3).

#### 4.8 Activité sur le marché du travail

Dans la présente section, on met de côté les rapports entre les caractéristiques démographiques de base des ménages et le genre de réponse, pour examiner les caractéristiques de l'activité sur le marché du travail. L'idée de faire cette évaluation est venue d'un besoin de mesurer le biais dû à la non-réponse qui peut se glisser dans les estimations de ces caractéristiques à partir des données de l'enquête.

La section 3.2 fait mention de différences considérables entre les répartitions des ménages répondants et non-répondants en fonction de la taille des ménages, alors que la section 3.1 décrit un phénomène semblable selon le nombre de mois d'inclusion dans l'échantillon. Pour cette raison, on présente



La répartition des répondants et des non-répondants en fonction du groupe d'âge et du nombre de mois qu'ils participent à l'enquête est présentée aux

#### 4.7 Âge des personnes et nombre de mois d'inclusion dans l'échantillon

Les tableaux 6a et 6b montrent que la non-réponse varie beaucoup selon la taille des ménages, mais que l'âge n'est pas un facteur important, bien qu'il existe un lien entre la taille des ménages et l'âge de leurs membres.

À dire ceux qui enregistrent le taux de non-réponse le plus grand. 65 ans et plus vivent dans des ménages formés d'une ou deux personnes, c'est-à-dire qu'il y a un fait (explicite plus haut) que la plupart des enquêtes de personnes âgées de 65 ans et plus est le plus élevé de tous les groupes d'âge. Ce phénomène est dû au fait (explicite plus haut) que la plupart des enquêtes de personnes âgées de 65 ans et plus enregistrent le taux de non-réponse le plus faible prise séparément (sauf les ménages de taille quatre en 1980), les enquêtes est peut-être le plus remarquable est que, pour chacune des tailles de ménage qu'on constate pour l'ensemble des ménages dans l'échantillon. Le fait que les taux de non-réponse et les groupes d'âge sont très différents de ceux augmentent. Par contre, pour une taille de ménage donnée, les rapports entre dire que les taux de non-réponse diminuent quand la taille des ménages diverses tailles de ménage, que celle observée dans la section 3.2, c'est-à-dire que les divers groupes d'âge affichent la même tendance pour les Les chiffres qui figurent aux tableaux 6a et 6b révèlent que les taux de non-

ménages composés d'une personne. spéciale dans les techniques appliquées pour compenser la non-réponse des dants et des non-répondants par groupe d'âge qui méritent une attention taille. Par conséquent, ce sont les écarts entre les répartitions des répon- de 5 % des enquêtes âgées de 25 à 44 ans vivent dans un ménage de cette 65 ans et plus font partie d'un ménage formé d'un seul membre, alors que moins ticulièrement importante, étant donné que près de 28 % des enquêtes âgées de 22.4 % comparativement à 34.3 % en 1980). Cette dernière observation est par-réponse que d'un ménage répondant (22.3 % comparativement à 34.3 % en 1981 et proportion beaucoup plus faible de personnes sont membres d'un ménage de non-

réponse les plus forts ont été constatés dans les groupes d'âge 65 ans et plus et 20-24 ans. Là encore, on remarque la relation inverse entre la taille du ménage et le taux de non-réponse. Or, les ménages comptant 1 ou 2 membres enregistrent les taux de non-réponse les plus élevés. Comme les personnes dans les groupes d'âge 65 ans et plus et 20-24 ans ont une probabilité plus grande de vivre seules ou en couple que les personnes des autres groupes, leurs taux de non-réponse devraient donc être élevés. La variation des taux de non-réponse des individus d'un groupe d'âge à l'autre indique un effet possible sur la qualité des estimations calculées à partir des données de l'enquête. Dans les groupes d'âge qui enregistrent un taux de non-réponse plus faible que le taux de l'ensemble des individus, l'importance des non-répondants sera surestimée si on utilise un facteur de correction de la pondération qui ne tient pas compte de la variable âge. Le phénomène inverse se produit lorsque le taux de non-réponse d'un groupe d'âge dépasse le taux de non-réponse de l'ensemble des personnes soumises à l'enquête. Jusqu'à un certain point, tout biais introduit au niveau provincial peut être compensé par la technique de correction du quotient.

#### 4.6 Âge des personnes et taille des ménages

On a également calculé la répartition des personnes par groupe d'âge et genre de réponse pour les diverses tailles de ménage. Ces répartitions, ainsi que les taux de non-réponse, figurent au tableau 6a pour 1980 et au tableau 6b pour 1981. Ces chiffres ont été obtenus à partir des moyennes annuelles des années étudiées.

Les répartitions des personnes par groupe d'âge chez les répondants et les non-répondants sont assez semblables dans le cas des ménages composés de 2, 3, 4 ou 5 personnes ou plus. Toutefois, parmi les ménages comptant un seul membre, on note des différences importantes entre la répartition par groupe d'âge des répondants et des non-répondants. Dans les ménages d'une personne, les principales différences se trouvent dans le groupe d'âge 25-44 ans qui compte beaucoup plus de personnes, selon la répartition, appartenant à des ménages non répondants qu'à des ménages répondants (39.6 % contre 28.8 % en 1981 et 35.5 % contre 27.9 % en 1980), et dans le groupe d'âge 65 ans et plus, où une



Le taux de non-réponse de toutes les personnes dans l'échantillon est de 3.13 %, en 1980, et de 2.63 % en 1981. Au niveau des ménages, ces taux s'élèvent à 4.02 % et à 3.43 % respectivement. Le fait que les taux de non-réponse individuels sont moins élevés que ceux des ménages confirme la relation inverse, signalée à la section 3.2, entre la taille du ménage et le taux de non-réponse. Comme les ménages nombreux ont un taux de non-réponse faible, une grande proportion des individus se classent dans la catégorie des répondants. Les relations observées entre les répartitions en pourcentage des répondants et des non-répondants individuels confirment les résultats présentés dans la section précédente relativement aux taux de non-réponse par âge 0-14 ans et 15-19 ans, les taux de non-réponse respectifs étaient de 2.50 % et 2.42 % en 1980, et de 2.12 % et 1.92 % en 1981. Les taux de non-

Bien que ce soit le ménage qui constitue généralement l'unité qui peut fournir des réponses, le tableau 5 présente la répartition en pourcentage des particuliers par groupe d'âge et genre de réponse. On trouve aussi dans ce tableau la répartition des non-répondants exprimée en pourcentage de l'ensemble de l'univers, ce qu'on pourrait appeler les taux de non-réponse au niveau des personnes visées par l'enquête.

#### 4.5 Age des personnes

d'estimations sur les familles. question est particulièrement importante en ce qui a trait à la production méthodes d'ajustement de la pondération appliquée à la non-réponse. Cette incidence de certains types de famille soit mal compensée dans les diverses avoir un certain effet sur le taux de non-réponse. Il est possible que l'in-Donc, outre la taille des ménages, la composition familiale des ménages semble pourcentages plus élevés de ménages non-répondants que de ménages répondants. (plus) et les ménages composés d'un couple marié seulement figurent pour des répondants. Les ménages formés uniquement de personnes seules (soit une ou nellement plus de ces types de ménage parmi les répondants que parmi les non-paraison des autres types de ménage. En d'autres termes, il y a proportion-désignées par les codes 6, 7 et 8, ont des taux de non-réponse faibles en com-

beaucoup plus grand de ménages composés d'une et deux personnes parmi les ménages non-répondants que parmi ceux qui répondent à l'enquête alors que, bien entendu, il y a moins de ménages de taille assez grande (3, 4 et 5 membres ou plus) parmi les non-répondants que parmi les répondants. Cette différence entre les répartitions devient encore plus importante lorsqu'on prend en compte le nombre de mois d'inclusion dans l'échantillon, ou le groupe de renouvellement, surtout dans le cas du premier et du deuxième mois. Après le deuxième mois, le taux de non-réponse a tendance à demeurer stable chez les ménages formés de deux personnes ou plus, alors que le taux de non-réponse des ménages qui comptent 1 ou 2 membres continue à varier jusqu'à la fin de leur inclusion dans l'échantillon. Ces observations portent à croire que la taille des ménages et le numéro de renouvellement sont des caractéristiques importantes dont il faut tenir compte dans l'évaluation des méthodes de correction de la non-réponse.

#### 4.4 Composition familiale des ménages

Dans la section 3.2, on a constaté des différences considérables dans la distribution des ménages répondants et non-répondants selon leur taille. Afin de poursuivre l'analyse des différences entre les ménages répondants et non-répondants, on a dressé un tableau qui présente des chiffres sur les types de famille qui forment les ménages. La composition familiale des ménages est basée sur le nombre de familles économiques à l'intérieur du ménage, la taille des unités familiales, la présence ou l'absence d'enfants et sur l'état matrimonial et l'âge du chef de l'unité familiale. Le tableau 4a indique les valeurs précises de ces variables qui correspondent à chaque type de famille, et le tableau 4b montre la répartition en pourcentage des ménages et le taux de non-réponse de chaque type de famille selon le genre de réponse.

Les taux de non-réponse très élevés des ménages formés d'un seul membre sont encore une fois évidents au tableau 4b. Ces taux sont particulièrement grands chez les ménages composés uniquement d'une personne seule âgée de moins de 65 ans. Les ménages qui comptent un couple marié ainsi que d'autres membres (qu'il s'agisse d'enfants ou d'adultes), c'est-à-dire les types de ménage

semblable à celui démontré par le tableau 2, où seule la taille du ménage était prise en considération. Le taux de non-réponse diminue lorsque la taille du ménage augmente. De même, le tableau 3 montre que, pour un nombre donné de mois d'inclusion dans l'échantillon (de un à six), le taux de non-réponse a tendance à diminuer à mesure que la taille du ménage augmente.

Compte tenu de ces deux tendances, le taux de non-réponse le plus élevé devrait apparaître parmi les ménages composés d'une personne, pendant le premier mois de participation à l'enquête. En outre, le taux de non-réponse le plus bas devrait être observé parmi les ménages qui comptent cinq personnes ou plus, au cours du dernier mois de participation à l'enquête (c'est-à-dire le sixième mois). D'après les moyennes annuelles calculées pour 1980 et 1981, cette supposition se confirme. Les taux de non-réponse les plus élevés, pour 1980 et 1981, sont de 13.39 % et de 12.81 % respectivement, et il s'agit de ménages composés d'une seule personne et qui n'ont pas répondu à l'enquête au cours du premier mois. En 1980, le taux de non-réponse le moins élevé était de 1.54 %; il a été obtenu des ménages formés de cinq personnes ou plus, au troisième mois d'enquête. Toutefois, un taux de 1.59 % a été observé chez les ménages composés de cinq personnes ou plus, au sixième mois. En 1981, le taux de non-réponse le plus faible (1.37 %) correspond aux ménages de cinq personnes ou plus, qui étaient à leur troisième mois dans l'échantillon. Le taux de non-réponse de ces ménages, au sixième mois, était de 1.39 %. Donc, même si le taux de non-réponse le plus bas ne se produit pas uniquement au dernier mois chez les ménages qui comptent cinq personnes ou plus, il n'y a pas de différence notable entre le taux observé dans cette catégorie de ménages et le taux qui est effectivement le plus faible.

La répartition de la taille des ménages en fonction du nombre de mois d'inclusion dans l'échantillon et du genre de réponse indique l'effet possible du biais de non-réponse sur les estimations. Une correction de la non-réponse qui ne tient pas compte de la taille des ménages permet de neutraliser l'effet des ménages non-répondants, si elle est fondée sur la répartition des ménages répondants. Autrement dit, une telle correction sous-estime l'importance des ménages comptant 1 ou 2 membres, mais surestime celle des ménages de 3 personnes ou plus. Si on examine les répartitions des ménages, on trouve un nombre



traitement des données de l'EPA, les ménages non-répondants sont représentés par des ménages qui, en moyenne, comptent un plus grand nombre de membres. On peut donc douter de la supposition que les ménages répondants représentent bien les ménages non-répondants, du moins en ce qui concerne la taille.

#### 4.3 Taille du ménage et nombre de mois d'inclusion dans l'échantillon

Dans les deux sections précédentes, on a souligné les variations importantes des taux de non-réponse établis en fonction du nombre de mois d'inclusion dans l'échantillon et de la taille du ménage. Le prochain tableau permet de voir si les différences entre les taux de non-réponse subsistent quand la taille du ménage ou le nombre de mois demeure constant. À partir des moyennes annuelles calculées pour 1980 et 1981, le tableau 3 présente la répartition en pourcentage des ménages répondants et non-répondants selon les catégories de taille de ménage et de nombre de mois de participation à l'enquête, de même que les taux de non-réponse pour 1980 et 1981.

On constate que, comme dans les tableaux 1 et 2, les tendances à la baisse des taux de non-réponse de l'ensemble des groupes de renouvellement sont aussi présentes lorsqu'une des variables est constante et que l'autre varie. Par exemple, au tableau 1, les taux de non-réponse pour toutes les tailles de ménage regroupées diminuent quand le nombre de mois d'inclusion dans l'échantillon augmente. Dans le tableau 3, le même phénomène se dégage du comportement des taux de non-réponse associés au nombre de mois, dans chaque catégorie de taille de ménage prise séparément. Comme lorsqu'on examine seulement la variable nombre de mois, on observe une diminution marquée du taux de non-réponse du premier au deuxième mois. En effet, pour chaque taille de ménage donnée, le taux de non-réponse régresse, au deuxième mois, d'environ la moitié de ce qu'il était au premier. Dans le cas des ménages de taille 1 et 2, le taux de non-réponse continue de baisser d'un mois à l'autre, alors que le taux de non-réponse des ménages composés de 3 personnes ou plus a tendance à demeurer stable après le deuxième mois.

Si on maintient le nombre de mois constant afin de voir le comportement du taux de non-réponse lorsque la taille du ménage varie, on obtient un résultat

Dans l'EPA, la non-réponse a généralement lieu au niveau des ménages, c'est-à-dire que le taux de non-réponse partielle à l'intérieur des ménages est très faible. Le ménage constitue l'unité où se produit la non-réponse. Il est donc nécessaire de caractériser les ménages pour mesurer les effets de la non-réponse sur les estimations calculées à partir de l'enquête - qu'elles correspondent au niveau des ménages, des familles ou d'unités individuelles. L'attribut du ménage qui est peut-être le plus fondamental pour la production d'estimations démographiques et socio-économiques dans l'enquête est la taille du ménage. Pour ce qui est de la collecte des données, il est raisonnable de supposer que la difficulté de communiquer avec des ménages diminue à mesure que la taille des ménages augmente.

Afin d'évaluer l'effet possible de la taille du ménage sur le taux de non-réponse, le tableau 2 présente la répartition en pourcentage des ménages en fonction de leur taille et du genre de réponse, d'après les moyennes obtenues pour les années civiles 1980 et 1981. Dans ces deux cas, le taux de non-réponse baisse considérablement lorsque la taille du ménage augmente. Le taux de non-réponse observé pour les diverses tailles de ménage varie, en 1980, d'un maximum de 7.48 % chez les ménages composés d'une personne à un minimum de 1.89 % chez ceux comprenant 5 personnes ou plus; les pourcentages respectifs pour l'année 1981 sont de 6.58 % et de 1.69 %. Un examen de la distribution des ménages répondants et non-répondants selon leur taille révèle des différences importantes. Si l'on examine ces deux répartition, on note que la proportion de ménages non-répondants composés d'une personne est presque deux fois celle des ménages répondants de même taille. Un peu plus de 50 % des ménages qui répondent à l'enquête comprennent 5 personnes ou plus, alors que seulement à peu près 30 % des ménages non-répondants ont cette taille. L'écart entre les répartition des ménages répondants et non-répondants selon la taille se traduit aussi par une différence entre la taille moyenne de ces ménages. En 1980, la taille moyenne des ménages répondants était de 2.93 personnes et celle des ménages non répondants était de 2.26, tandis qu'en 1981 les tailles moyennes respectives étaient de 2.88 et de 2.19. Il semble donc que, dans la correction de la non-réponse effectuée à l'étape du



Pour cette raison, on a examiné la composition de l'échantillon en fonction du nombre de mois de participation à l'enquête et du genre de réponse. Des estimations pondérées du nombre de ménages, à l'échelle du pays, selon le nombre de mois dans l'échantillon et le genre de réponse ont été calculées à partir des moyennes obtenues pour 1980 et 1981 (voir le tableau 1). Étant donné l'attention accordée, lors de la conception du plan de sondage, au caractère représentatif de l'échantillon, on s'attendrait à voir une répartition égale du total des valeurs pondérées en fonction du nombre de mois dans l'échantillon. En effet, les données révèlent que tout près d'un sixième (ou 16.67 %) de l'ensemble des ménages sont regroupés dans chaque catégorie du nombre de mois de participation à l'enquête. Dans tous les cas, la différence entre les distributions en pourcentage à l'intérieur de chacune des catégories est de moins de un demi de 1 %.

Lorsqu'on examine la distribution des ménages en fonction du nombre de mois d'inclusion dans l'échantillon et du genre de réponse, on observe un éloignement d'une répartition uniforme, surtout parmi les non-répondants. Cette constatation est illustrée par le rapport entre les taux de non-réponse et le nombre de mois dans l'échantillon. Comme le démontre le tableau 1, le taux de non-réponse diminue à mesure que le nombre de mois de participation à l'enquête augmente. La baisse la plus forte a lieu entre le premier et le deuxième mois, alors que le taux de non-réponse correspond à environ la moitié de ce qu'il était au premier mois. On note d'autres diminutions de ce taux pour les mois qui suivent. Entre le deuxième et le sixième mois, le taux de non-réponse subit une baisse de 21.1 % et de 34.2 % respectivement pour 1980 et 1981.

On observe une tendance semblable à la baisse dans la répartition en pourcentage des ménages non-répondants à mesure que le nombre de mois d'inclusion dans l'échantillon augmente. Dans la répartition en pourcentage, il y a beaucoup plus de ménages non-répondants, au premier mois, que de ménages répondants. Toutefois, ce pourcentage diminue en fonction de la durée d'inclusion dans l'échantillon. Ainsi, toute estimation calculée pour un groupe de renouvellement avec une correction de la non-réponse établie pour l'ensemble de ces groupes peut produire des estimations légèrement biaisées.

Comme il a été expliqué dans l'introduction, l'EPA est fondée sur un plan de renouvellement des unités de l'échantillon, chaque groupe d'unités demeurant dans l'échantillon pendant une période de six mois. A l'étape de l'élaboration du plan d'échantillonnage, on prend bien soin de s'assurer que le groupe correspondant à chaque numéro de renouvellement (c'est-à-dire les logements qui forment chaque groupe) est un sous-échantillon représentatif constitué d'un système de l'échantillon global de l'EPA. Dans le passé, on a souligné le phénomène du biais attribuable au renouvellement, c'est-à-dire que la valeur prévue des estimations calculées à partir d'un seul groupe de renouvellement varie en fonction du nombre de mois d'inclusion dans l'échantillon.

#### 4.1 Nombre de mois de participation à l'enquête

La méthode exposée dans la section précédente décrit les techniques d'estimation des totaux de caractéristiques à partir du fichier longitudinal requis pour cette étude. Dans la présente section, on examine les différences pour un certain nombre de variables (séparément et conjointement) entre les unités de réponse et les unités de non-réponse. Chacune des sections suivantes porte sur une variable en particulier ou sur un classement recoupé de variables. On y trouve aussi une explication de l'intérêt d'examiner ces variables, ainsi que des tableaux où figurent des totalisations des variables analysées et un résumé des principales conclusions.

### 4. ANALYSE

La méthode de pondération fasse varier les estimations d'une étude à l'autre. Dans le cas présent, les dossiers individuels sont pondérés au moyen d'un produit de la fraction de sondage inverse, du facteur de sous-pondération de grappe et d'un facteur de compensation<sup>5</sup>. Dans l'examen et l'interprétation des résultats de la section 3, ou dans une comparaison de ces résultats avec ceux obtenus dans une autre étude de la non-réponse, il est important de se rappeler que la source des données est le fichier de données longitudinales, que seuls les ménages qui ont participé à l'enquête pendant au moins un mois entrent dans le calcul des estimations, et que la structure de la pondération repose seulement sur des facteurs de pondération du plan d'échantillonnage.

Mois de	non-réponse	imputer les données manquantes
1	2, 3, 4, 5, 6	2, 3, 4, 5, 6
2	1, 3, 4, 5, 6	1, 3, 4, 5, 6
3	2, 4, 1, 5, 6	2, 4, 1, 5, 6
4	3, 5, 2, 6, 1	3, 5, 2, 6, 1
5	4, 6, 3, 2, 1	4, 6, 3, 2, 1
6	5, 4, 3, 2, 1	5, 4, 3, 2, 1

Advenant qu'il n'y a pas de mois de réponse, aucune imputation n'est effectuée et le ménage en question est exclu de la présente étude.

### 3.4 Mise en garde

Si on compare les taux de non-réponse établis dans la présente étude avec ceux obtenus pour les groupes de renouvellement de l'échantillon mensuel de l'EPA, on constatera des différences de grandeur. La principale cause de cet écart est l'exclusion de certains ménages non répondants de cette étude sur les données longitudinales. Tel qu'il a été expliqué dans une section précédente, on est en mesure de caractériser un ménage pendant un mois de non-réponse seulement s'il existe des données relatives à ce ménage pour un autre mois. En d'autres termes, il doit y avoir au moins un mois de réponse pour qu'un ménage non répondant puisse être caractérisé. Par conséquent, un ménage qui ne répond pas à l'enquête ou qui correspond à une combinaison d'un ménage non-répondant et d'un logement vacant pendant les six mois où il fait partie de l'échantillon de l'enquête est exclu de la présente étude. Ainsi, certains ménages non-répondants qui entrent dans le calcul mensuel du taux de non-réponse de l'EPA ne sont pas pris en compte dans l'analyse longitudinale du taux de non-réponse présentée ici. Environ 1.4 % de l'ensemble des ménages échantillonnés ont été exclus pour cette raison.

L'exclusion de certains ménages non répondants soit la cause principale des différences entre les résultats de la présente étude et ceux de n'importe quelle autre étude sur la non-réponse basée sur les données mensuelles de l'EPA. Outre cette cause de différences entre les résultats, il se peut que



Quant aux ménages qui fournissent des réponses au moins une fois au cours des six mois pendant lesquels ils sont inclus dans l'échantillon, les renseignements donnés pendant les mois de réponse compléteront ce qui manque pour les mois de non-réponse. De cette façon, on peut estimer les caractéristiques des ménages de non-répondants. Pour appliquer cette méthode d'imputation, il faut que les ménages concernés aient participé à l'enquête à au moins une occasion; il se peut également que cette participation s'étende sur plusieurs mois. Dans ce dernier cas, le mois de réponse le plus rapproché du mois de non-réponse servira à compléter les données manquantes. Si deux mois de réponse sont séparés d'un mois de non-réponse par le même intervalle, les renseignements du mois précédant le mois de non-réponse seront choisis pour imputer les données incomplètes. Le tableau suivant résume cette technique.

Pour déterminer les caractéristiques des ménages répondants, on examine les caractéristiques de chaque membre du ménage qui a répondu à l'enquête. Toutefois, dans le cas des ménages non-répondants, il faut appliquer une technique d'imputation. Les caractéristiques de ces ménages devraient être identiques aux caractéristiques des membres du ménage qui ont répondu à l'enquête pendant un mois donné ou être approximativement représentées par les caractéristiques de ces membres.

mois donné, il y a trois genres de réponse: les répondants, les non-répondants et les logements vacants. Les répondants sont les ménages qui ont rempli le questionnaire de l'EPA pour tous les membres du ménage admissibles à l'enquête ou pour quelques-unes de ces personnes. Les non-répondants sont des ménages composés de personnes qui devraient faire partie de l'enquête mais qui, pour une raison ou pour une autre, décident de ne pas participer ou ne sont pas en mesure de participer à cause de certaines circonstances. Par contre, les logements vacants sont des logements inoccupés ou occupés par des personnes non incluses dans l'univers de l'enquête. Les logements identifiés comme vacants ne sont donc pas pris en compte dans l'étude des caractéristiques des ménages répondants et non-répondants.

répondants au cours de la période où ils sont soumis à l'enquête, on peut observer des aspects dynamiques de l'activité sur le marché du travail. Dans un mois donné, les logements qui composent un des six groupes de renouvellement arrivent à la fin de la période de participation à l'enquête. Il est possible d'analyser la composition de chaque ménage à l'intérieur de ce groupe, ainsi que l'évolution des réponses, pendant les cinq mois antérieurs. Ce genre d'observation peut se faire à l'aide du fichier de données longitudinales qui fournit un enchaînement des renseignements recueillis auprès d'un ménage donné durant les six mois où il a fait partie de l'enquête.

Dans l'EPA, on attribue un code de désignation unique à chaque logement et à chaque personne, ce qui permet de relier les renseignements sur les personnes, les ménages et les logements pour les six mois où un ménage est inclus dans l'enquête, et de constituer ainsi le fichier de données longitudinales.

D'abord, des enregistrements longitudinaux contenant les codes-réponses des six mois d'enquête sont créés pour chaque logement. Si le ménage qui occupe un logement figure dans la catégorie des répondants pendant un mois ou plus, on inclut aussi des enregistrements individuels de renseignements sur les membres du ménage qui habitaient ce logement au moment où les réponses ont été données. Cependant, si aucune réponse n'a été fournie pendant les six mois, on possède alors seulement des renseignements de base sur le logement. Toute personne qui était membre d'un ménage à un moment donné durant les six mois d'enquête est associée à un enregistrement dans le fichier de données longitudinales et, à partir de cet enregistrement, il est possible d'obtenir des renseignements concernant des variables démographiques ou l'activité sur le marché du travail pour les mois où un membre d'un ménage était un répondant. La structure des données longitudinales permet d'examiner les ménages qui répondent et les ménages qui ne répondent pas, et d'évaluer les caractéristiques de chaque genre de réponse.

### 3.3 Méthode d'estimation

Dans l'étude des caractéristiques des ménages répondants et non-répondants, il faut connaître le genre de réponse de tous les ménages à chaque mois. Pour un



quoï ces logements sont remplacés par un autre groupe de logements, de sorte que, tous les mois, un sixième de l'échantillon est remplacé ou renouvelé. Ainsi, pendant un mois donné, l'échantillon de l'EPA contient six groupes de logements et chaque groupe se trouve à une étape différente du processus de renouvellement. C'est-à-dire qu'un groupe fait partie de l'enquête pour la première fois (le nouveau groupe de renouvellement), un autre pour la deuxième fois, ..., et le dernier pour la sixième fois.

A chaque mois pendant une semaine, la semaine d'enquête<sup>1</sup>, les interviewers de l'EPA communiquent avec les résidents des logements sélectionnés afin d'obtenir des renseignements concernant la composition du ménage, des variables démographiques et l'activité sur le marché du travail des membres du ménage compris dans l'univers de l'enquête<sup>2</sup>. Pour diverses raisons, les interviewers ne peuvent pas toujours recueillir des renseignements sur tous les logements sélectionnés. Ces logements, où aucune interview n'a lieu, sont classés "Logements vacants" ou comme étant occupés par des ménages de non-répondants<sup>3</sup>, selon le mode d'occupation. Dans le cas des logements vacants, il est impossible d'obtenir une réponse et aucune réponse n'est prévue, tandis que, dans le cas des ménages non-répondants, il manque des données d'enquête. Une correction<sup>4</sup> de la non-réponse pour compenser les données manquantes est effectuée lors du traitement des données, suivant l'hypothèse que les ménages qui ont été interviewés, c'est-à-dire les unités de réponse, représentent bien les ménages qui auraient dû être interviewés, c'est-à-dire les unités de non-réponse. Si cette hypothèse est fautive, la correction de la non-réponse introduit alors un biais dans les estimations de l'enquête. Ce biais augmente à mesure que le taux de non-réponse s'accroît. Pour cette raison, il est important que les caractéristiques des ménages qui répondent et de ceux qui ne répondent pas à l'enquête soient semblables, et on fait de grands efforts pour minimiser la non-réponse.

### 3.2 Fichier de données longitudinales

Les estimations calculées à partir des données mensuelles transversales de l'EPA fournissent une image statique de la population et du marché du travail pour chaque mois. Toutefois, en appariant les renseignements fournis par les

Le pourcentage de personnes occupées diminue à mesure que le nombre de mois d'inclusion augmente, tandis que le pourcentage de chômeurs subit une hausse. Quant aux ménages qui comptent deux membres ou plus, les différences entre les répartitions des ménages répondants et non répondants en fonction de l'activité sont beaucoup moins prononcées que dans le cas des ménages formés d'une personne, mais le pourcentage de chômeurs parmi les ménages composés de deux personnes ou plus s'accroît généralement à mesure que le nombre de mois d'inclusion dans l'échantillon augmente.

Il pourrait y avoir des avantages à appliquer au processus de correction de la non-réponse des variables autres que la taille des ménages et le nombre de mois d'inclusion dans l'échantillon, qui concernent l'activité sur le marché du travail. On pourrait alors améliorer les estimations de la population active. Toutefois, il se peut que le besoin d'un ajustement global des poids, de même que la petite taille des échantillons à ce degré d'agrégation et le niveau relativement faible de la non-réponse à l'heure actuelle dans l'EPA ne permettent pas de faire une correction de la non-réponse basée sur des variables reliées à l'activité sur le marché du travail. Une correction de la non-réponse qui utilisait la taille du ménage et le nombre de mois d'inclusion devrait pourtant améliorer les estimations de la population active. Par conséquent, il pourrait être utile de considérer des corrections qui se servent de deux groupes de ménages, ceux qui comptent un seul membre et ceux qui comptent deux membres ou plus, ainsi que des corrections basées sur deux valeurs du nombre de mois d'inclusion, un mois et deux mois ou plus. Cette étude devrait permettre d'évaluer la possibilité d'améliorer les techniques actuelles de correction de la non-réponse dans l'EPA.

### 3. SOURCE DE DONNÉES

#### 3.1 L'enquête sur la population active

L'EPA est fondée sur un plan d'échantillonnage aléatoire stratifié à plusieurs degrés, et la stratification est faite en fonction des régions économiques de chaque province. Le logement constitue l'unité finale d'échantillonnage. Les logements sélectionnés demeurent dans l'échantillon pendant six mois, après

renouvellement à l'aide d'une correction de la non-réponse établie pour l'en-  
semble de ces groupes produise des estimations légèrement biaisées.

Dans le cas de la variable taille des ménages, la non-réponse diminue à mesure  
que la taille des ménages augmente. Si l'on examine les répartitions des  
ménages répondants et non répondants selon leur taille, on note que la propor-  
tion de ménages non répondants composés d'une personne est presque deux fois  
celle des ménages répondants de même taille et que la proportion de ménages  
répondants formés de 5 personnes ou plus est plus que le double de celle des  
ménages non répondants. Il semble donc qu'une correction de la non-réponse  
qui ne tient pas compte de la taille des ménages comporte généralement une  
représentation des ménages de non-réponse par des ménages qui comptent un plus  
grand nombre de membres que les ménages de non-réponse.

La répartition des ménages répondants et non répondants selon la taille du  
ménage et le nombre de mois d'inclusion dans l'échantillon demeure inchangée,  
lorsqu'on examine ces deux caractéristiques ensemble. Étant donné que l'ana-  
lyse de ces deux variables révèle une relation fonctionnelle étroite avec le  
taux de non-réponse, une correction de la non-réponse qui tient compte de la  
taille des ménages et du nombre de mois d'inclusion dans l'échantillon devrait  
aider beaucoup à réduire le biais de renouvellement dans les estimations des  
ménages, des familles économiques et des caractéristiques dépendantes de ces  
variables.

On a constaté que, outre la taille des ménages et le nombre de mois d'inclu-  
sion dans l'échantillon, l'activité sur le marché du travail, en particulier  
le chômage, était également liée au genre de réponse. Parmi les ménages non  
répondants, il y a un accroissement du pourcentage de chômeurs quand le nombre  
de mois de participation de ces ménages à l'enquête grandit mais, dans le cas  
des ménages répondants, ce pourcentage est relativement stable. Lorsqu'on  
prend en compte la taille du ménage, on observe une relation bien définie pour  
les ménages composés d'une personne, et une relation un peu plus variable pour  
les ménages formés de deux membres ou plus. Chez les ménages comptant un  
membre, les proportions de personnes occupées et de chômeurs sont beaucoup  
plus élevées parmi les non-répondants que parmi les répondants. Par ailleurs,



L'échantillon, la région, l'âge des membres du ménage et leur activité. Cette étude est basée sur des données tirées des fichiers de données longitudinales de l'EPA. La section 3 fournit une brève description de l'EPA, des fichiers longitudinaux et des méthodes utilisées pour caractériser les ménages de non-réponse. Ensuite, la section 4 présente les données calculées à partir des fichiers longitudinaux et en donne une analyse. La dernière section décrit brièvement l'importance des résultats de la présente étude en ce qui concerne la qualité des données de l'EPA au niveau des particuliers, des familles et des ménages. On y propose également des méthodes possibles pour venir à bout du problème de la non-réponse, de façon à compenser ou à minimiser les lacunes dans les données d'enquête causées par ce problème.

## 2. RÉSUMÉ DES PRINCIPALES CONCLUSIONS

Dans l'EPA, les méthodes de compensation de la non-réponse reposent sur l'hypothèse selon laquelle les caractéristiques des ménages non répondants sont semblables à celles des ménages répondants. Advenant le cas où cette supposition est fautive, la technique de correction de la non-réponse ajoutera un biais aux estimations de l'enquête. Il est impossible d'évaluer avec exactitude l'importance du biais de non-réponse mais, par un examen des données longitudinales concernant toute la période d'inclusion d'un ménage dans l'échantillon de l'enquête, on peut tracer un profil des ménages répondants et non-répondants, et mesurer l'ampleur des différences entre ces groupes.

Parmi les diverses variables examinées dans la caractérisation des ménages répondants et non répondants, celles relatives au nombre de mois d'inclusion dans l'échantillon, à la taille du ménage et à l'activité sur le marché du travail affichent un lien bien défini avec le genre de réponse. Dans le cas de la variable nombre de mois d'inclusion dans l'échantillon, on observe que les taux de non-réponse diminuent à mesure que le nombre de mois augmente. Entre le premier et le deuxième mois, le pourcentage de ménages de non-réponse accuse une très forte baisse, pour ensuite diminuer progressivement jusqu'au sixième mois, ce qui porte à croire que le nombre de mois d'inclusion dans l'échantillon est un facteur déterminant en ce qui concerne le genre de réponse. Il se peut donc que les estimations calculées pour un groupe de

# CARACTÉRISTIQUES DES MÉNAGES RÉPONDANTS ET NON RÉPONDANTS DANS L'ENQUÊTE SUR LA POPULATION ACTIVE DU CANADA

Elizabeth Clayton Paul, Murray Lawes<sup>1</sup>

Cet article présente les résultats d'une étude qui a été faite pour caractériser les ménages répondants et non-répondants dans l'EPA. Cette étude a été motivée par deux projets associés au remaniement de l'EPA, c'est-à-dire l'estimation des familles et l'évaluation des procédures pour corriger la non-réponse. Toutefois, les résultats de cette étude ont aussi un intérêt plus général vis-à-vis de l'évaluation de la qualité des données qui proviennent de l'EPA.

## 1. INTRODUCTION

La non-réponse est l'absence de renseignements complets sur toutes les unités sélectionnées pour une enquête par échantillonnage ou pour un recensement. Elle pose des problèmes particulièrement difficiles aux responsables d'une enquête et aux utilisateurs des données. Essentiellement, la non-réponse agit sur la qualité des données d'une enquête de deux façons. D'abord, elle diminue la taille de l'échantillon, ce qui entraîne une perte de précision des estimations de l'enquête. Deuxièmement, si les différences entre les unités de réponse et les unités de non-réponse ne sont pas bien prises en compte dans les méthodes d'estimation, la non-réponse peut introduire un biais dans les estimations de l'enquête. La présente étude porte sur ce dernier aspect de la qualité des données, notamment la caractérisation des unités de réponse et de non-réponse dans l'enquête sur la population active du Canada (EPA). Les pages qui suivent donnent une idée des effets possibles de la non-réponse sur les estimations d'une enquête et proposent quelques variables dont il faut tenir compte dans la modifications faites pour corriger la non-réponse. Les unités sont définies en fonction des variables taille du ménage, type de famille économique, la période pendant laquelle l'unité fait partie de

<sup>1</sup> Elizabeth Clayton Paul, Groupe des caractéristiques économiques, Statistique Canada et Murray Lawes, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.



- [19] Schabie, W.L. (1979), "A Composite Estimator for Small Area Statistics", In Synthetic Estimates for Small Areas (J. Steinberg, Ed.) National Institute on Drug Abuse Research Monograph n° 24, U.S. Government Printing Office, Washington, D.C., 36-53.
- [20] Singh, M.P. et Tessier, R. (1975), "Some Estimators for Domain Totals", Journal of The American Statistical Association 71, 322-325.
- [21] Sonquist, J.N. et Morgan J.A. (1964), "The Detection of Interaction Effects", Monograph n° 35, Survey Research Center, Institute for Social Research, University of Michigan.

[10] National Center for Health Statistics (1968), "Synthetic State Estimates of Disability", P.H.S. Publication No. 1759, U.S. Government Printing Office, Washington, D.C.

[11] Platek, R. et Singh, M.P. (1976), Méthodologie de l'enquête sur la population active du Canada, n° 71-526 au catalogue, Statistique Canada.

[12] Purcell, N.J. et Linacre, S. (1976), "Techniques for the Estimation of Small Area Characteristics", document présenté à la 3<sup>e</sup> conférence des statisticiens australiens, Melbourne, Australie.

[13] Purcell, N.J. et Kish, L. (1979), "Estimation for Small Domains", *Biometrics* 35, 365-384.

[14] Royall, R.M. (1973), "Discussion of two papers on Recent Developments in Estimation of Local Areas", *Proceedings of the American Statistical Association*, Social Statistics Section, 43-44.

[15] Royall, R.M. (1978), "Prediction models in Small Area Estimation", NIDA Workshop on Synthetic Estimates, Princeton, N.J.

[16] Sarndal, C.E. (1981), "When Robust Estimation is not an obvious answer: The case of the Synthetic Estimator versus Alternatives for Small Areas", *Proceedings of the American Statistical Association*, Survey Research Section.

[17] Schabie, W.L., Brock, D.B. et Schnack, G.A. (1977), "An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics", *Proceedings of the American Statistical Association*, Social Statistics Section, 1017-1021.

[18] Schabie, W.L. (1978), "Choosing Weights for Composite Estimators for Small Area Statistics", *Proceedings of the American Statistical Association*, Survey Research Section, 741-746.

# BIBLIOGRAPHIE

- [1] Drew, J.D. et Choudhry, G.H. (1979), "Small Area Estimation", rapport technique, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada.
- [2] Changuarde, P.D. et Singh, M.P. (1976), "Synthetic estimation in the LFS", rapport technique, Division de l'élaboration d'enquêtes-ménages, Statistique Canada.
- [3] Changuarde, P.D. et Singh, M.P. (1977), "Synthetic Estimates in periodic household surveys", Techniques d'enquête, Statistique Canada, 3, 152-181.
- [4] Changuarde, P.D. et Singh, M.P. (1978), "Evaluation of Efficiency of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 53-61.
- [5] Gonzalez, M.E. (1973), "Use and Evaluation of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 33-36.
- [6] Gonzalez, M.E. et Waksberg, J. (1973), "Estimation of the Error of Synthetic Estimates", document présenté à la première réunion de l'International Association of Survey Statisticians, Vienne, Autriche.
- [7] Gonzalez, M.E. (1975), "Small Area Estimation of Unemployment", Proceedings of the American Statistical Association, Social Statistics Section, 437-460.
- [8] Gonzalez, M.E. et Hoza, C. (1978), "Small Area Estimation with application to Unemployment and Housing Estimates", Journal of the American Statistical Association 73, 7-15.
- [9] Holt, T., Smith, T.M.F. et Tomberlin, T.J. (1979), "A Model Based Approach to Estimation for Small Sub-groups of a Population", Journal of the American Statistical Association, 74, 405-410.

Tableau 5 Degré de dépendance moyen de l'estimateur dépendant de l'échantillon séparé vis-à-vis de la composante synthétique: divisions de recensement de la Nouvelle-Écosse

Division de recensement	Dépendance (1- $\delta$ ) vis-à-vis de la composante synthétique		Proportion de la population de la division de recensement	
	$K_0=0.5$	$K_0=1.0$	Strates partielles	Strates complètes
201	.04	.15	1.00	-
202	.15	.20	.70	.30
203	.01	.12	1.00	-
204	.12	.22	1.00	-
205	.04	.14	1.00	-
206	.03	.08	.37	.63
207	.05	.07	.26	.74
210	.06	.09	.35	.65
211	.04	.05	.13	.87
212	.06	.10	.52	.48
213	.03	.16	1.00	-
214	.04	.16	1.00	-
215	.11	.21	1.00	-
216	.05	.15	1.00	-
217	.01	.01	.03	.97
218	.18	.28	1.00	-

Tableau 4 CEF pour lesquelles le biais de l'estimateur synthétique séparé pour les chômeurs dépasse 10 %

<u>Catégorie (i)</u>			<u>Catégorie (ii)</u>			<u>Catégorie (iii)</u>		
<u>Biais rel. %</u>			<u>Biais rel. %</u>			<u>Biais rel. %</u>		
<u>Données CEF</u>	<u>à jour</u>	<u>Données pas à jour</u>	<u>Données CEF</u>	<u>à jour</u>	<u>Données pas à jour</u>	<u>Données CEF</u>	<u>à jour</u>	<u>Données pas à jour</u>
102	12.25	-3.90	414	1.57	17.78	301	12.71	25.85
104	-12.52	6.23	426	-7.38	-11.78	304	-11.29	-17.57
411	-13.10	-0.37	450	1.93	-11.35	412	-10.07	-15.80
436	10.48	-0.48	455	2.67	15.46	438	12.15	14.22
474	18.35	-6.04	460	-7.74	20.80	501	10.52	16.83
806	15.15	-0.20	504	7.19	41.05	579	15.50	10.59
			527	9.59	11.76	818	10.83	39.41
			605	3.63	14.29			
			701	4.90	17.74			
			804	0.78	15.85			
			813	-0.48	-15.46			

Catégorie (i): le biais dépasse 10 % seulement lorsque l'information auxiliaire est à jour.

(ii): le biais dépasse 10 % seulement lorsque l'information auxiliaire n'est pas à jour.

(iii): le biais dépasse 10 % dans les deux cas.



Tableau 3 Biais absolu relatif moyen (%) des estimateurs synthétiques pour les CEF (l'information auxiliaire comprenant des données pas à jour sur la population des 15 ans et plus)

Caractéristique	Personnes occupées		Chômeurs	
	Séparé	Combiné	Séparé	Combiné
<u>Niveau de construction</u>				
<u>Province</u>				
Terre-Neuve	1.20	2.19	1.75	3.17
Ile-du-Prince-Édouard	3.54	5.62	3.71	2.80
Nouvelle-Écosse	0.87	1.64	1.25	1.87
Nouveau-Brunswick	2.95	2.54	6.35	7.04
Québec	2.53	3.87	3.87	4.51
Ontario	2.17	3.52	3.01	4.25
Manitoba	1.42	2.28	2.22	3.44
Saskatchewan	2.41	2.65	4.35	4.85
Alberta	5.24	7.08	5.12	10.33
Colombie-Britannique	1.73	2.71	2.72	4.20

Tableau 2 Efficacité des estimateurs pour les petites régions par rapport à l'estimateur direct -  
divisions de recensement de la Nouvelle-Écosse (données auxiliaires pas à jour)

Caractéristique	Variable auxiliaire	Niveau de construction	ESTIMATEUR		Dépendant de l'échantillon ( $K_0=1.0$ )
			Pour domaines stratifiés a posteriori	Synthétique	
Personnes occupées	Population	combiné	3.47	4.73	4.07
	Population par âge/sexe	"	3.60	4.73	4.23
	Population	combiné	1.44	2.19	1.68
Chômeurs	Population par âge/sexe	"	1.46	2.21	1.69

Tableau 1 Efficacité des estimateurs pour les petites régions par rapport à l'estimateur direct - divisions de recensement de la Nouvelle-Écosse (données auxiliaires à jour)

ESTIMATEUR							
Caractéristique	Variable auxiliaire	Niveau de construction	Pour domaines stratifiés a posteriori		Composite ( $\alpha^* \neq 0.223$ )	Dépendant de l'échantillon	
			Synthétique			$K_0=0.5$	$K_0=1.0$
Personnes occupées	Logement	combiné	4.58	10.17	10.92	9.17	10.42
	Population	"	4.92	10.75	10.58	10.50	11.67
	Population par âge/sexe	"	5.08	10.83	11.25	11.17	12.25
	Logement	séparé	2.75	10.50	-	9.58	10.50
	Population	"	2.83	10.92	-	10.58	11.42
	Population par âge/sexe	"	2.83	11.00	-	11.00	11.75
Chômeurs	Logement	combiné	1.33	1.70	1.75	1.40	1.55
	Population	"	1.36	1.70	1.75	1.43	1.58
	Population par âge/sexe	"	1.36	1.70	1.75	1.43	1.58
	Logement	séparé	1.30	1.69	-	1.48	1.58
	Population	"	1.33	1.69	-	1.51	1.61
	Population par âge/sexe	"	1.33	1.69	-	1.51	1.61

relatives à des domaines de la taille des CEF ou des divisions de recensement soit d'utiliser des techniques d'estimation perfectionnées (et d'effectuer une agrégation des estimations sur une certaine période), au moyen de données du recensement et de données d'enquête seulement. Toutefois, pour pouvoir satisfaire ces besoins à plus long terme et pour répondre à la demande d'autres types de données basées sur des enquêtes de moins grande envergure et portant sur d'autres types et tailles de domaines, il faudra utiliser largement les trois sources de données que sont le recensement, les enquêtes et les dossiers administratifs. Des estimateurs de régression linéaire à plusieurs variables du type envisagé par Erickson (1974) et par Gonzalez et Hoza (1978), qui font usage de ces trois sources, devraient faire l'objet d'une étude détaillée afin d'en déterminer le biais, l'erreur quadratique moyenne et les difficultés de calcul. Chacune de ces sources possède ses propres limites, mais elles offrent ensemble un potentiel considérable d'amélioration, c'est-à-dire que les points faibles d'une source peuvent être les points forts d'une autre. On peut donc envisager avec optimisme la possibilité que des techniques statistiquement sûres exploitant de façon harmonisée les points forts de données de diverses sources permettent d'obtenir dans l'avenir des renseignements de bonne qualité sur les petites régions pour un large éventail de sujets.

#### REMERCIEMENT

Des consultations avec M. R. Platek se sont révélées utiles au moment de la rédaction finale du présent document.

Au Canada, pour des caractéristiques qui font l'objet d'enquêtes régulières à grande échelle (comme l'enquête sur la population active), il semble que la meilleure façon de satisfaire dans l'immédiat à la demande de données

Les estimateurs sans transformation de la structure, proposés par Purcell et Kish (1980), sont un autre sujet de recherche prometteur. Dans ce cas, le processus d'estimation, défini par la structure d'association (c.-à-d. la relation entre les variables  $y$  et  $x$  à un moment antérieur à l'échelle du domaine) et la structure de répartition (c.-à-d. la relation initiale présente dans l'échelle de la grande région), maintient la relation initiale présente dans la structure d'association tout en laissant intacte l'information actuelle fournie dans la structure de répartition.

En ce qui concerne l'enquête sur la population active, puisque les méthodes d'estimation pour les petites régions appliquées à des domaines non prévus se sont révélées supérieures aux estimations non biaisées fondées sur le plan de sondage pour des domaines prévus comparables, il pourrait être souhaitable d'étendre les travaux de recherche à certains petits domaines prévus (type a). En particulier, l'estimateur dépendant de l'échantillon examiné ici, ainsi que des estimateurs semblables analysés par d'autres auteurs feront l'objet de recherches complémentaires sur les caractéristiques de la population active. En outre, ces travaux devraient s'appliquer à d'autres enquêtes de moins grande envergure de Statistique Canada pour lesquelles on a besoin de données à l'échelle des petites régions. Il faudra en outre travailler davantage à l'élaboration de méthodes d'estimation de la variance pouvant être appliquées à ces estimateurs.

utilisant uniquement des données du recensement et des données d'enquête, sur-tout pour des domaines non prévus (type c). Les estimateurs analysés sont fondés de diverses façons sur un estimateur synthétique et un estimateur pour domaines stratifiés a posteriori, l'objectif étant de trouver le juste milieu entre le biais et l'erreur quadratique moyenne. Sont indiquées ci-dessous certaines des voies que pourraient emprunter les recherches futures en vue de l'élaboration de techniques statistiquement sûres de production de données à l'échelle des petites régions au Canada.



4. L'estimateur composite construit sous forme de combinaison linéaire de l'estimateur pour domaines stratifiés a posteriori et de l'estimateur synthétique est plus efficace que l'un ou l'autre des estimateurs qui le composent, bien que la différence soit non significative par rapport à la composante synthétique, pour une valeur optimum de  $\alpha$ . Son biais dépend du poids attribué à la composante synthétique, puisque le biais de l'estimateur pour domaines stratifiés a posteriori est en général négligeable. En outre, comme le calcul de la valeur optimum de  $\alpha$  est très complexe, on ne peut en pratique utiliser qu'une valeur estimative de  $\alpha$ , d'où l'efficacité réduite de cet estimateur.

5. Les estimateurs synthétique, composite et dépendant de l'échantillon présentent tous une efficacité plus ou moins égale, lorsque  $K_0 = 1$ , et sont supérieurs à l'estimateur non biaisé pour les domaines prévus basé sur le plan de sondage.

6. Puisque le biais de l'estimateur synthétique séparé est inférieur à celui de l'estimateur synthétique combiné, la version séparée de l'estimateur dépendant de l'échantillon présente un biais relatif moindre que celui de sa version combinée. Le biais de la composante pour domaines stratifiés a posteriori séparée peut être contrôlé par la combinaison des strates dont l'intersection avec le domaine est très petite. Par conséquent, si on tient compte des trois facteurs que sont le biais, l'erreur quadratique moyenne et les complexités de calcul, c'est l'estimateur dépendant de l'échantillon construit au niveau des strates et utilisant la population selon l'âge et le sexe qui semble être le meilleur choix. L'attribution d'une valeur à  $K_1$  dépend aussi de plusieurs facteurs. Dans le cas présent, la valeur  $K_1 = 1$  s'est révélée efficace tout en assurant une protection contre le biais de l'estimateur synthétique.

## 5. ORIENTATION DES RECHERCHES FUTURES

L'étude dont fait état le présent document portait sur l'évaluation de certaines méthodes d'estimation pour les petites régions, dans le contexte de l'EPA,

Comme on le prévoyait également, les valeurs moyennes de  $(1-\delta)$ , lorsque  $K_0 = 0,5$ , sont inférieures à celles obtenues lorsque  $K_0 = 1,0$ , ce qui veut dire que plus la valeur de  $K_0$  choisie est faible, plus la valeur de  $(1-\delta)$  est basse, de sorte que l'estimateur dépendant de l'échantillon présente une moins grande dépendance (poids) vis-à-vis de la composante synthétique. Toutefois, comme il est illustré au tableau 1, il faut faire un compromis entre le biais et l'efficacité, puisque des valeurs moins élevées de  $K_0$  entraînent également une réduction de l'efficacité. Les valeurs de  $K_0$  mentionnées ci-dessus ont permis d'obtenir un degré de confiance raisonnable pour le type de domaines qui nous intéresse ici, mais il est possible en général de choisir d'autres valeurs de  $K_0$  en fonction, par exemple, de la taille du domaine, de la taille de l'échantillon, de la taille des strates et de leur délimitation par rapport au domaine.

#### 4.5 Conclusion

1. La variable population selon l'âge et le sexe donne de meilleurs résultats de façon constante que toute autre variable auxiliaire, quoique sa supériorité par rapport à la variable population totale des 15 ans et plus soit peu significative.

2. L'estimateur pour domaines stratifiés a posteriori, bien qu'il soit plus efficace que l'estimateur pour domaines simples, présente des faiblesses par rapport aux trois autres estimateurs pour les petites régions qui ont été analysés.

3. L'estimateur synthétique séparé produit un biais relatif inférieur à celui de l'estimateur synthétique combiné. En outre, les biais moyens ont tendance à être plutôt faibles et à ne s'accroître que légèrement lorsque l'information auxiliaire n'est plus à jour, alors que les biais pour des domaines particuliers peuvent être très élevés et présenter des variations considérables, ce qui rend vains les efforts en vue d'identifier les "valeurs extrêmes" lorsqu'on recherche une moins grande dépendance par rapport aux estimateurs synthétiques.

que l'estimation finale comporte un "niveau de confiance raisonnable". Il est également important de déterminer la dépendance vis-à-vis de l'estimateur synthétique sans rendre trop complexe le traitement informatique. Compte tenu de ces facteurs, il y a lieu, dans le contexte de l'enquête sur la population active, de rechercher des estimateurs pour les petites régions dont le rendement au niveau des domaines non prévus est comparable à celui d'estimation simples de l'enquête pour des domaines prévus. Une fois connus les estimateurs qui satisfont à ce critère, il faudrait mettre l'accent davantage sur la réduction du biais que sur l'amélioration de l'efficacité, en particulier si les différences d'efficacité sont minimes.

Les variances moyennes de l'estimateur non biaisé de l'enquête pour les domaines prévus (disons  $\bar{X}$ ) de taille comparable aux domaines non prévus ont été obtenues de manière analogue à la moyenne des eqm définie en (3.1). L'efficacité des estimateurs synthétique, composite et dépendant de l'échantillon par rapport à l'estimation normale de l'enquête pour les domaines prévus, c.-à-d.  $X$ , a également été calculée. Elle varie de 1,08 à 1,17 pour les chômeurs, et de 1,22 à 1,47 pour les personnes occupées, de sorte que les trois estimateurs satisfont au critère mentionné plus haut. Comme l'estimateur dépendant de l'échantillon s'en remet à l'estimateur synthétique dès qu'il n'y a pas un échantillon "suffisant" dans le domaine, son biais dépend du poids lié à la composante estimateur synthétique, poids sur lequel on peut agir par un choix approprié de  $K_0$ . Au tableau 5 sont présentées les valeurs de (1- $\delta$ ), en moyenne pour 100 répétitions, avec  $K_0 = 0,5$  et  $K_0 = 1,0$ , pour la version séparée de l'estimateur dépendant de l'échantillon, la variable auxiliaire étant la population totale des 15 ans et plus pour chacune des divisions de recensement (domaines non prévus) de la présente étude. Ces valeurs moyennes de (1- $\delta$ ) indiquent le degré de dépendance de l'estimateur dépendant de l'échantillon vis-à-vis de la composante synthétique. Comme il fallait s'y attendre, les domaines constitués surtout de strates partielles ont tendance à dépendre davantage de la composante synthétique, bien que dans une très faible mesure. Par exemple, si  $K_0 = 1$ , la valeur la plus élevée est 0,28 pour la division de recensement 218.

L'estimateur synthétique affiche en général un biais élevé, tout en étant très efficace. Il faut donc se demander, dans la recherche d'un estimateur raisonnable pour les petites régions, dans quelle mesure on peut réduire l'effet du biais de l'estimateur synthétique sans trop sacrifier son efficacité, de façon

#### 4.4 Choix de l'estimateur - compromis entre l'efficacité et le biais

tercensitaire.

décèles lorsqu'on établit des estimations courantes au cours de la période in-  
les cas correspondant à la catégorie (ii) du tableau 4, qui ne peuvent être  
Toutefois, le risque de biais de l'estimateur synthétique existe toujours dans  
pourrait dans de tels cas établir la valeur de  $K_0$  à un niveau moins élevé.  
exemple, en ce qui a trait à l'estimateur dépendant de l'échantillon, on  
biais élevé à l'époque à laquelle remonte l'information auxiliaire. Par  
l'estimateur synthétique lorsqu'on sait que, pour un domaine, il produisait un  
Cette constatation semble indiquer qu'il faudrait faire un usage moindre de  
10 % pour la période postérieure où cette information n'était pas à jour.  
dépassait 10 % avec l'information auxiliaire à jour, il excédait également  
lorsqu'elle ne l'était pas. De plus, dans la moitié des cas où le biais  
(sur 279) respectivement lorsque l'information auxiliaire était à jour et  
toutefois, l'examen du tableau 4 montre qu'il a dépassé 10 % dans 13 et 18 CEF  
Le biais de l'estimateur synthétique a été en moyenne relativement peu élevé,

des tendances analogues à celles présentées au tableau 3.

1971. Bien qu'ils se soient révélés légèrement inférieurs, ces biais suivent  
formation auxiliaire étaient l'une et l'autre basées sur le recensement de  
On a également calculé les biais dans le cas où la variable étudiée et l'in-

thèse d'homogénéité.

en résulte en général est important, en raison de l'affaiblissement de l'hypo-  
niveau où on construit l'estimateur synthétique est élevé, plus le biais qui  
confirme ce qu'on pouvait prévoir intuitivement, c'est-à-dire que plus le  
tique combiné, pour les deux caractéristiques étudiées. Cette constatation  
l'estimateur synthétique séparé est moindre que celui de l'estimateur synthé-



#### 4.2. Etude de l'efficacité - information auxiliaire pas à jour

Dans cette partie de l'étude, le plan de sondage et l'information auxiliaire étaient basés sur les résultats du recensement de 1971, tandis que la variable étudiée était fondée sur les données de 1976. Comme on peut le constater au tableau 2, bien que dans le cas des chômeurs l'utilisation de techniques d'estimation pour les petites régions ait produit, par rapport à l'estimateur direct, des résultats supérieurs à ceux obtenus dans le cas où l'information auxiliaire était à jour, cette supériorité a été considérablement moindre dans le cas des personnes occupées, ce qui s'explique probablement par la corrélation moins grande entre la variable étudiée et l'information auxiliaire lorsque le plan de sondage et l'information auxiliaire ne sont tous deux plus à jour. On observe également dans ce cas que l'efficacité de l'estimateur synthétique est supérieure pour chacune des deux caractéristiques mesurées.

#### 4.3 Etude du biais

Comme le biais de l'estimateur pour domaines stratifié a posteriori est généralement négligeable, le biais de l'estimateur composite et celui de l'estimateur dépendant de l'échantillon sont en général inférieurs à celui de l'estimateur synthétique, c.-à-d. qu'ils ne proviennent que du degré de dépendance de l'estimateur vis-à-vis de la composante synthétique. On a donc étudié en détail le biais de l'estimateur synthétique. La population totale des 15 ans et plus étant utilisée comme variable auxiliaire, les biais relatifs pour les caractéristiques personnes occupées et chômeurs ont été calculés et sont donnés au tableau 3 pour les dix provinces. Ces biais se rapportent au cas où les domaines non prévus sont les circonscriptions électo- rales fédérales et où les variables étudiées sont fondées sur les données du recensement de 1976, tandis que le plan de sondage et les facteurs d'ajuste- ment (poids synthétiques) sont basés sur le recensement de 1971. On a également calculé les biais en utilisant comme variable auxiliaire des sous- groupes selon l'âge et le sexe et on a découvert que, tout en étant légèrement moins élevés, ils suivaient des tendances analogues. L'examen de ce tableau révèle qu'à l'exception des deux plus petites provinces, soit l'I.-P.-É. (pour les chômeurs) et le N.-B. (pour les personnes occupées), le biais relatif de



version combinée est environ deux fois plus efficace que la version séparée, ce qui est probablement attribuable à l'effet plus accentué sur cette dernière du regroupement en grappes dans le plan de sondage.

Comme l'estimateur pour domaines stratifié a posteriori était moins efficace dans sa forme séparée, on s'attendait à ce qu'il en soit de même pour l'estimateur composite, et on n'a donc étudié que la forme combinée de ce dernier. Par ailleurs, on a découvert que la forme séparée de l'estimateur dépendant de l'échantillon était fonction un peu plus de la composante synthétique, de sorte que le niveau de construction n'a pas influé sur l'efficacité.

ii) Effet de l'information auxiliaire À titre de variable auxiliaire, la population selon l'âge et le sexe s'est révélée uniformément supérieure, bien que de façon peu significative, à la population totale des 15 ans et plus, dans le cas des quatre estimateurs utilisant de l'information auxiliaire. En outre, ces deux variables sont de bien meilleures variables auxiliaires que les chiffres de logements.

Pour l'enquête proprement dite, le choix de la population selon l'âge et le sexe comme variable auxiliaire peut également être souhaitable en prévision de la correction des estimations pour tenir compte des biais dus à la non-réponse et au sous-dénombrement, puisque ces deux facteurs peuvent dépendre de l'âge et du sexe.

iii) Comparaison entre les estimateurs Dans le cas des chômeurs, l'estimateur composite avec  $\chi^2$  optimum choisi pour la caractéristique chômeurs se révèle supérieur, mais de façon non significative, aux autres estimateurs peu importe le niveau de construction, et le choix de la variable auxiliaire n'a en apparence aucun effet notable sur aucun des estimateurs. La situation n'est pas aussi claire dans le cas des personnes occupées, mais l'estimateur dépendant de l'échantillon démontre une légère supériorité par rapport aux autres estimateurs, en particulier lorsque la variable auxiliaire est la population selon l'âge et le sexe.

et

$$B = \sum ( \sum X \quad \sum Y \quad \sum h \tilde{h} a h g ) - \frac{(\sum h \tilde{h} a h g)^2}{\sum Y}$$

où  $a_{hg}$  et  $Y_{hg}$  sont définis au chapitre 2, et où  $X_{hg}$  et  $a_{hg}$  sont définis dans les mêmes termes pour la variable  $x$  (sur la base du recensement). On a calculé les biais absolus relatifs au niveau des provinces en faisant la somme des biais absolus pour les différentes CEF et en la divisant par le total provincial de la variable  $x$ .

#### 4. ANALYSES DES RÉSULTATS

##### 4.1 Etude de l'efficacité - information auxiliaire à jour

Pour cette partie de l'étude empirique (Monte-Carlo), les données qui ont servi à la simulation du plan de sondage et les variables auxiliaires utilisées pour l'estimation portent sur la même période que la variable étudiée, c.-à-d. celle du recensement de 1971. L'efficacité des quatre estimateurs pour les petites régions par rapport à l'estimateur direct est illustrée au tableau 1, pour les niveaux de construction séparés et combinés, et pour chacune des variables auxiliaires suivantes: logements, population totale (15 ans et plus), et population selon les groupes d'âge et le sexe. Les divisions de recensement de la Nouvelle-Écosse comptant de 3 885 à 39 260 habitants ont été utilisées comme domaines non prévus (type c) pour les besoins de l'étude. Les observations suivantes peuvent être formulées:

##### i) Comparaison entre un estimateur séparé et un estimateur combiné

Le niveau de construction des estimateurs n'influe pas sensiblement sur l'efficacité des estimateurs synthétiques, tant pour les personnes occupées que pour les chômeurs. Dans le cas de l'estimateur pour domaines stratifié a posteriori appliqué aux personnes occupées, cependant, la

Nous ne fournissons les résultats que pour les 16 divisions de recensement de la Nouvelle-Écosse. On a obtenu des résultats analogues pour les autres domaines non prévus qui ont été analysés.

Soit  $\hat{x}_{m(r)}$  l'estimation du total  $x$  (c.-à-d. le total de la variable  $x$  dans le domaine  $a$ ), pour la  $r^{\text{ème}}$  répétition, dans le cas de la méthode  $m$  d'estimation pour les petites régions. La moyenne des erreurs quadratiques moyennes calculée selon la méthode  $m$  pour les 16 domaines de l'étude a été obtenue suivant la formule suivante:

$$\text{moy eqm (m)} = \frac{1}{16} \sum_{r=1}^{16} (x - x_m(r))^2 / 100 \quad (3.1)$$

L'efficacité de l'estimateur pour les petites régions (m) par rapport à l'estimateur direct, par exemple la méthode  $m_0$ , est donnée par:

$$\text{Eff (m vs } m_0) = \frac{\text{Moy eqm (m)}}{\text{Moy eqm (m}_0)} \quad (3.2)$$

#### 3.4 Évaluation du biais des estimateurs synthétiques

Comme on connaissait pour tout le Canada la composition de la base de l'EPA et les circonscriptions électorales fédérales aussi bien pour les unités de recensement de 1971 que pour celles de 1976, il a été possible de calculer les biais exacts des estimateurs synthétiques au moyen des données du recensement. Les cas suivants ont été étudiés: (i) plan de sondage et information auxiliaire à jour (auquel cas le plan de sondage, les facteurs d'ajustement et les variables  $x$  étaient tous basés sur le recensement de 1971); et (ii) plan de sondage et information auxiliaire pas à jour (auquel cas le plan de sondage et les facteurs d'ajustement étaient basés sur le recensement de 1971, tandis que les variables  $x$  étaient basées sur le recensement de 1976). Soit  $a_{BS}$  et  $a_{CS}$  les biais respectifs des estimateurs synthétiques séparé et combiné pour le domaine non prévu  $a$ , on obtient alors:

$$B_{aSS} = \sum g h \tilde{h} - \frac{\sum a h g}{\sum X} - \frac{\sum a h g}{\sum X} \quad (3.3)$$

personnes occupées et les inactifs, les sous-groupes suivants comptaient pour environ 25 % de la variation: (1) Les personnes de 15-16 ans et de 65 ans et plus; (ii) Les femmes de 17 à 64 ans; (iii) Les hommes de 17 à 64 ans. On n'a obtenu aucune amélioration appréciable en divisant davantage ces sous-groupes. Outre les estimateurs basés sur les sous-groupes énumérés ci-dessus, on a examiné les estimateurs fondés sur la population totale des 15 ans et plus, et sur les chiffres de logements auxquels ont été pris en compte parce qu'il est possible d'obtenir entre les recensements des données à jour sur ce sujet et qui sont désagrégées au niveau requis. Il est à remarquer que les estimateurs qui utilisent ces chiffres de population (15 ans et plus) et de logements ne sont que des cas spéciaux de la formule générale où le nombre de sous-groupes de la population est égal à 1.

### 3.3 Évaluation de l'efficacité des estimateurs pour les petites régions

L'application de la technique de Monte-Carlo a porté sur 16 divisions de recensement (DR) et 11 circonscriptions électorales fédérales (CEF) de la Nouvelle-Écosse, ainsi que sur 7 CEF d'ailleurs au Canada. (La Nouvelle-Écosse comprend en tout 18 DR, mais deux ont été éliminées parce qu'elles correspondaient à des strates complètes de l'EPA.) Étant donné qu'il s'agit d'un plan de sondage à plusieurs degrés et que l'étude portait sur un grand nombre de domaines, le coût du traitement informatique a été élevé et on a décidé d'utiliser seulement 100 répétitions.

Il convient de souligner que les divisions de recensement et les circonscriptions électorales fédérales constituent respectivement des réseaux d'unités géostatistiques et géopolitiques couvrant le Canada. Elles sont les unes et les autres au nombre d'environ 300. La population des circonscriptions électorales fédérales est relativement uniforme et se situe entre 80 000 et 120 000 habitants, tandis que celle des divisions de recensement, lesquelles correspondent souvent à des régions administratives locales ou à des comtés, varie considérablement.



Pour les variables état matrimonial, âge et sexe, on a appliqué la méthode de détection automatique des interactions (DAI), mise au point par Sonquist et Morgan (1964), à un échantillon de données du recensement provenant de partout au Canada afin d'obtenir des sous-groupes optimaux de la population, c'est-à-dire un pour chaque caractéristique de la population active. Les résultats de l'analyse DAI ont révélé que, pour les chômeurs, aucun sous-groupe de la population n'intervenait pour plus de 2 % de la variation, tandis que pour les

Les estimateurs définis au chapitre 2 exigent une information auxiliaire pour constituer des sous-groupes de la population. Comme l'EPA n'est remaniée qu'une fois tous les dix ans, il serait souhaitable que les sous-groupes de la population puissent être basés non seulement sur les données du recensement mais aussi sur les données du recensement quinquennal, de façon que l'information auxiliaire puisse être mise à jour à mi-chemin du cycle de l'enquête. Des variables comme l'activité ou la profession ont donc été exclues; il restait comme choix possibles pour les sous-groupes de la population divers-  
ses classifications recoupées de variables démographiques de base.

## 3.2 Choix des sous-groupes de population

Pour le cas (ii), on a limité l'étude aux UNAR. Les unités primaires et secondaires ont été choisies à l'aide des chiffres du recensement de 1971, mais l'échantillon de personnes à l'intérieur des unités secondaires était basé sur les résultats du recensement de 1976. L'information auxiliaire a été tirée des données du recensement de 1971.

Dans les SD ruraux et les centres urbains, les deux degrés finals de l'échantillonnage ont été simulés au moyen d'échantillons aléatoires de personnes. Dans le cas des UNAR, les SD formant les strates géographiques étaient connus, mais pour le reste le plan de sondage de l'EPA était indépendant du recensement. Pour les besoins de l'étude, on a divisé au hasard les SD en "grappes" dont la distribution des tailles correspondait à celle des grappes de l'EPA. Pour chaque répétition, on a prélevé un échantillon de "grappes" et un échantillon aléatoire de personnes à l'intérieur de ces grappes.



### 3. DESCRIPTION DE L'ÉTUDE EMPIRIQUE

#### 3.1 Simulation du plan de sondage de l'EPA

L'EPA est fondée sur un plan de sondage aréolaire à plusieurs degrés (voir Platek et Singh, 1976). À l'intérieur de chacune des dix provinces du Canada, deux types principaux d'unités sont établis: les unités autorensementaires (UAR) qui correspondent en général à des villes de 15 000 habitants et plus, et les unités non autorensementaires (UNAR) qui correspondent à de petits centres urbains et à des régions rurales. Les villes qui constituent les UAR sont divisées en strates géographiques compactes comptant chacune 15 000 habitants, à l'intérieur desquelles on prélève un échantillon à deux degrés de grappes (semblables à des îlots) et de logements.

Dans le cas des UNAR, les régions économiques, dont le nombre peut atteindre 10 par province, constituent le point de départ. On délimite dans chacune de 1 à 5 strates de 30 000 à 80 000 habitants, à partir des données du recensement sur sept grandes classes d'activités économiques. On subdivise ensuite les strates en unités primaires d'échantillonnage (UPÉ) comprenant de 2 000 à 5 000 habitants. Dans les parties rurales des UPÉ, comprenant de 2 000 à 5 000 habitants. Dans les dénominations (SD) du recensement de 1971, qui comptent approximativement 500 habitants; dans les parties urbaines, tous les centres urbains sont choisis avec certitude. Les deux derniers degrés correspondent aux grappes et aux logements.

Deux cas ont été examinés aux fins de la simulation du plan de sondage de l'EPA: (i) le cas où le plan de sondage et l'information auxiliaire sont à jour, et (ii) le cas où ils ne sont pas à jour.

Pour le cas (i), le plan de sondage, l'information auxiliaire et les variables à l'étude étaient fondés sur les données du recensement de 1971. Les chiffres de population (15 ans et plus) classés par recoupement selon l'âge et le sexe, et la situation vis-à-vis de l'activité ont été extraits au niveau des SD. Dans le cas des UNAR, pour chaque répétition utilisée dans l'étude de simulation (Monte-Carlo), on a prélevé des échantillons indépendantes d'unités primaires et d'unités secondaires à partir des chiffres de population ou de

où

$$\delta = 1, \text{ si } \frac{\sum y}{\sum y} \geq K$$

$$= \frac{1}{K} \frac{\sum y}{\sum y}, \text{ autrement.}$$

Les rapports

$$\frac{\sum y}{\sum y} \text{ et } \frac{\sum y}{\sum y}$$

indiquent si, pour un échantillon donné, les sous-groupes de la population dans chaque strate ou dans un domaine sont trop ou pas assez représentés pour ce qui a trait aux renseignements auxiliaires de la variable "y". Lorsque la valeur de ces rapports est supérieure ou égale à 1, les sous-groupes de la population pour la variable auxiliaire y dans l'échantillon donné sont aussi bien ou mieux représentés que si un échantillon indépendant avait été prélevé pour le domaine selon la même fraction que pour la strate.

Une valeur convenable de  $K_0$  peut être attribuée. Dans la présente étude, on vérifie l'efficacité des estimateurs dépendants de l'échantillon pour deux valeurs précises de  $K_0$ , à savoir 1,0 et 0,5.

Holt, Smith et Tomberlin (1979) ont trouvé, en appliquant des méthodes de prévision, un estimateur (qui est une fonction des estimations synthétiques et directes) dont le poids affecté à la composante directe varie uniquement en fonction de la taille de l'échantillon compris dans un domaine. Sarnadal (1981), pour sa part, a proposé un estimateur différent dont le poids attaché à la composante directe dépend du rapport entre la taille de l'échantillon situé dans un domaine et la taille de l'échantillon contenu dans la grande région.

L'estimateur dépendant de l'échantillon (Drew et Choudhry, 1979), qui représente un cas particulier de l'estimateur composite, dépend du résultat de l'échantillon et se calcule assez facilement. Sa construction est influencée par le rendement de l'estimateur pour domaines stratifiés a posteriori, qui varie en fonction de la proportion de l'échantillon regroupée dans un domaine. Lorsque cette proportion est assez grande, l'estimateur dépendant de l'échantillon est identique à l'estimateur pour domaines stratifiés a posteriori, mais autrement, il se transforme en estimateur composite soumis à une dépendance progressivement plus forte vis-à-vis de l'estimateur synthétique (c'est-à-dire que son poids augmente) à mesure que la taille de l'échantillon dans le domaine diminue. Ainsi, l'estimateur séparé qui dépend de l'échantillon (construit au niveau de chaque strate) est défini par l'équation suivante:

$$\hat{\chi} = \sum \delta \left[ \frac{a_{hg}}{W} + (1 - \delta) \frac{a_{hg}}{W} \right] \quad (2.12)$$

où

$$\delta_{hg} = 1, \text{ si } \frac{a_{hg}}{W} \geq K_0, \text{ autrement, } \delta_{hg} = \frac{1}{K_0} \frac{a_{hg}}{W}$$

De même, l'estimateur combiné qui dépend de l'échantillon (construit au niveau d'un domaine) s'écrit:

$$\hat{\chi} = \sum \delta \left[ \frac{a_{hg}}{W} + (1 - \delta) \frac{a_{hg}}{W} \right] \quad (2.13)$$

et

$$\hat{X}_{sc}^a = \alpha_1 \hat{X}_{sp}^a + (1 - \alpha_1) \hat{X}_{ss}^a \quad (2.8)$$

$$\hat{X}_{cc}^a = \alpha_2 \hat{X}_{cp}^a + (1 - \alpha_2) \hat{X}_{cs}^a \quad (2.9)$$

Les valeurs optimales de  $\alpha_1$  et de  $\alpha_2$  pour réduire l'eqm au minimum sont données par

$$\alpha_1^* = \frac{\text{eqm} [\hat{X}_{ss}^a] - E [\hat{X}_{ss}^a] [\hat{X}_{sp}^a - aX] + \text{eqm} [\hat{X}_{ss}^a] + \text{eqm} [\hat{X}_{sp}^a] - 2 E [\hat{X}_{ss}^a - aX] [\hat{X}_{sp}^a - aX]}{\text{eqm} [\hat{X}_{ss}^a] - E [\hat{X}_{ss}^a] [\hat{X}_{sp}^a - aX] + \text{eqm} [\hat{X}_{ss}^a] + \text{eqm} [\hat{X}_{sp}^a] - 2 E [\hat{X}_{ss}^a - aX] [\hat{X}_{sp}^a - aX]} \quad (2.10)$$

et par une expression semblable pour  $\alpha_2^*$

Or, si on ne tient pas compte du terme de covariance dans la formule (2.10) en supposant que ce terme est petit par rapport à l'eqm  $[\hat{X}_{ss}^a]$  et l'eqm  $[\hat{X}_{sp}^a]$ , on obtient l'approximation suivante du poids optimal  $\alpha_1$ :

$$\alpha_1^{**} = \frac{\text{eqm} [\hat{X}_{ss}^a]}{\text{eqm} [\hat{X}_{ss}^a] + \text{eqm} [\hat{X}_{sp}^a]} \quad (2.11)$$

et une équation semblable pour  $\alpha_2^{**}$ , ce qui correspond à la manière dont Schabile (1978) a défini les coefficients de pondération.

## 2.4 Estimateurs dépendant de l'échantillon

En pratique, on ne connaîtra pas la vraie valeur de  $\alpha_1^*$  (ou de  $\alpha_2^*$ ) qui sert de poids dans l'équation de l'estimateur composite, parce qu'elle fait intervenir la variance et la covariance de la population, éléments qu'il faut estimer à partir de l'échantillon. Un calcul plus poussé du terme de covariance dans l'expression (2.10) pourrait être assez complexe et il faudrait donc peut-être utiliser la valeur approximative  $\alpha_1^{**}$  (ou  $\alpha_2^{**}$ ) qui requiert seulement les eqm estimées des deux estimateurs constitutifs de l'estimateur composite ou le rapport estimé de leur eqm. Dans un cas ou dans l'autre, ces estimations provoquent une certaine instabilité du poids utilisé, ce qui modifie le rendement de l'estimateur composite.

seulement les strates qui font partie du domaine, ce qui, croyait-on, fait diminuer le biais. Mais, en général, il n'est pas essentiel de limiter h de cette manière et on pourrait inclure d'autres régions voisines si elles sont censées satisfaire l'hypothèse d'homogénéité. Le biais et l'erreur quadratique moyenne (eqm) de ce genre d'estimateur ont été analysés par quelques-uns des auteurs mentionnés plus haut.

## 2.3 Estimateurs composites

Un estimateur composite dont les deux composantes sont un estimateur direct et un estimateur synthétique a été proposé par Royall (1973) et par d'autres auteurs, et Schabie (1978) en a fait une analyse. Ce genre d'estimateur réduit au minimum la probabilité de situations extrêmes (en ce qui concerne le biais et l'erreur quadratique moyenne) et on pourrait donc le trouver préférable à l'une ou à l'autre de ses composantes. La variance des estimateurs synthétiques est basse, puisqu'on leur applique les données d'une grande région pour calculer des estimations relatives à une petite région (un petit domaine) mais, pour la même raison, un biais assez grand peut se glisser dans les résultats si, comme on l'a mentionné précédemment, l'hypothèse d'homogénéité n'est pas satisfaite. Par contre, l'estimateur pour domaines simples, qui est sans biais, peut avoir une grande variance, surtout si la partie de l'échantillon qui se trouve dans le domaine est très petite. Les résultats de procédés empiriques démontrant le rendement relatif des estimateurs synthétiques et directs sont présentés par Gonzalez et Waksberg (1975), Schabie, Brock et Schnack (1977), et Ghangurde et Singh (1977). On obtient l'estimateur composite, que nous examinons dans cette section, en remplaçant l'estimateur direct (2.3) par l'estimateur pour domaines stratifiés a posteriori, qui peut avoir un léger biais mais qui est généralement plus efficace que l'estimateur direct.

Les deux types d'estimateur composite, c'est-à-dire le type séparé et le type combiné, sont formés par combinaison linéaire du type correspondant d'estimateur synthétique et d'estimateur pour domaines stratifiés a posteriori. Bref,



$$\bar{X}_{cs} = \frac{\sum h \tilde{a}_{hg}}{\sum h \tilde{a}_{hg}} = \frac{\sum h \tilde{a}_{hg}}{\sum h \tilde{a}_{hg}} \quad (2.7)$$

où  $t_{hg}$  correspond au total de la variable  $x$  dans l'échantillon pour le sous-groupe  $g$  de la population dans la strate  $h$ .

Ces estimateurs ont été étudiés par Purcell et Linacre (1976) et aussi par Ghanugde et Singh (1976, 1977, 1978), qui ont formulé des expressions pour la variance et le biais et évalué les estimateurs synthétiques avec des données de recensement et un modèle de superpopulation. Une forme différente d'estimateur synthétique avait été proposée par le National Centre for Health Statistics (1968) et examinée par Gonzalez (1973), Gonzalez et Waksberg (1973) et Gonzalez et Hoza (1975, 1978) à partir de données de la Current Population Survey aux États-Unis.

La différence entre un estimateur synthétique et un estimateur pour domaines stratifiés a posteriori devient évidente lorsqu'on compare les équations (2.4) et (2.6). L'estimateur pour domaines stratifiés a posteriori ne sert que de la partie de l'échantillon regroupée dans un domaine (c'est-à-dire  $t_{hg}$ ), tandis que le rapport des vraies valeurs aux valeurs estimées pour la variable  $y$  constitue le facteur d'ajustement qui, par conséquent, peut être inférieur ou supérieur à 1 (son espérance mathématique étant de 1). Par contre, l'estimateur synthétique utilise l'estimation de la variable  $x$  obtenue pour l'ensemble d'une strate partiellement recouverte par un domaine (notamment  $W_h \cdot t_{hg}$  pour  $h\tilde{a}$ ) et cette valeur est ensuite dégonflée par un facteur d'ajustement propre à chaque sous-groupe de la population (c.-à-d. le rapport de la variable  $y$  pour le domaine à la variable  $y$  pour la strate entière).

L'estimateur synthétique sera entaché d'un biais selon que se confirme plus ou moins l'hypothèse que la variable  $x$  est homogène entre un domaine et la grande région,  $\tilde{h}$ , à l'intérieur des sous-groupes de la variable  $y$ . Dans la définition d'un estimateur synthétique donnée plus haut, la grande région comprenait

Ainsi, l'estimateur séparé pour les domaines stratifié a posteriori (où les ajustements sont effectués au niveau de chaque strate) s'écrit:

$$\hat{X}_{sp} = \frac{\sum_y g h \tilde{y}}{\sum_y a h} \cdot t \quad (2.4)$$

où  $t_{hg}$  est le total de la variable  $x$  dans l'échantillon pour le sous-groupe  $g$  de la population à l'intersection du domaine "a" et de la strate  $h$ .

Quant à l'estimateur combiné pour les domaines stratifié a posteriori (où les ajustements sont effectués au niveau du domaine), il a la forme suivante:

$$\hat{X}_{cp} = \frac{\sum_y g h \tilde{y}}{\sum_y h \tilde{y} a h} \cdot t \quad (2.5)$$

L'estimateur pour domaines stratifié a posteriori est sans biais, abstraction faite du biais de l'estimation par quotient, à condition que  $a_{hg}$  et  $a_{hg}$  soient obtenus en même temps et de la même source comme, par exemple, d'un recensement.

Des estimateurs de ce type ont déjà fait l'objet d'une analyse par Singh et Tessier (1976), mais les strates établies a posteriori n'étaient pas du même genre.

## 2.2 Estimateurs synthétiques

Les estimateurs synthétiques séparé et combiné se définissent respectivement comme suit:

$$\hat{X}_{ss} = \frac{\sum_y g h \tilde{y}}{\sum_y a h} \cdot t \quad (2.6)$$

distinguer les estimateurs séparés et combinés selon le niveau d'application de l'ajustement. On peut écrire ces estimateurs sous la forme  $\hat{X}_{uv}^a$ , où est le niveau d'application de l'ajustement dont les valeurs possibles sont:

u = s : séparé  
= c : combiné

et v indique le type d'estimateur:

v = p : pour domaines stratifié a posteriori  
= s : synthétique  
= c : composite  
= d : dépendant de l'échantillon

Par exemple,  $\hat{X}_{cs}^a$  décrit un estimateur combiné synthétique, et ainsi de suite.

## 2.1 Estimateur pour domaines stratifié a posteriori

Définissons

$Y_{hg}^a$  = le total de la variable auxiliaire y pour le sous-groupe g dans la strate h, et

$Y_{hg}^a$  = le total de la variable auxiliaire y pour le sous-groupe g en  $a_h$ .

De plus, supposons que  $\hat{Y}_{hg}^a$  est une estimation sans biais de  $Y_{hg}^a$  et qu'on l'obtient de façon semblable à l'estimation directe définie dans l'équation (2.1), sauf que la caractéristique à estimer est la variable auxiliaire y dont on connaît la valeur pour l'ensemble des unités de l'échantillon (s) à un degré du plan d'échantillonnage (alors que (2.1) est définie pour la variable x dans l'échantillon des unités finales). En fait, s'il existe des renseignements sur la variable auxiliaire y, on peut se servir des unités d'échantillonnage à n'importe quel degré jusqu'à l'avant-dernier.

$a^t_h$  = total de la variable x dans  $a_h$

pour  $h=1, 2, \dots, L$ . Notons que  $a^t_h = 0$  pour tout  $h \notin \tilde{h}$ .

Ainsi, l'estimateur direct (aussi appelé l'estimateur fondé sur le plan de sondage ou l'estimateur simple pour domaines) du total de la variable x pour les unités dans "a", que nous exprimons par  $a^t_h$ , provient de l'équation:

$$\hat{X}^a = \sum_h W_h \cdot a^t_h \cdot \tilde{h} \quad (2.3)$$

Il convient de souligner que l'estimateur direct (2.3) ne se sert d'aucun renseignement auxiliaire; il faut simplement repérer toutes les unités de l'échantillon qui appartiennent à chaque domaine. À cause de l'échantillonnage par grappes, il se peut que le nombre d'unités dans un domaine donné soit très petit ou nul, ce qui entraîne une variance élevée pour cet estimateur.

Les autres estimateurs dont il sera question dans cette section utilisent, de diverses manières, des renseignements auxiliaires sur une variable y qui représente souvent le nombre de personnes dans les sous-groupes de la population (établis par répartition selon l'âge et le sexe, etc.) dans un recensement récent. Ces estimateurs sont des types suivants:

- 1) pour domaines stratifiés a posteriori
- 2) synthétique
- 3) composite
- 4) dépendant de l'échantillon

En outre, les estimateurs 2) à 4) sont influencés dans différentes mesures par la partie de l'échantillon à l'extérieur d'un domaine.

Pour chacun de ces estimateurs, les ajustements faits à l'aide des données auxiliaires peuvent correspondre à des ajustements séparés appliqués à chaque strate dont une partie est comprise dans un domaine donné, ou à un ajustement général appliqué à toutes les strates qui recouvrent un domaine. On peut donc

## 2. DESCRIPTION DE LA METHODE D'ESTIMATION

Considérons une population finie composée de N unités (par exemple, des ménages ou des logements dans les enquêtes-ménages), divisées en L strates désignées 1, 2, ..., h, ..., L. La stratification a été faite en fonction de certaines caractéristiques géographiques et (ou) socio-économiques, et la répartition de l'échantillon assure une certaine précision des estimations relatives aux strates. Le problème posé consiste à estimer le total d'une variable x pour toutes les unités qui appartiennent à un domaine aréolaire non prévu (type c). Nous représentons par "a" l'ensemble des unités comprises dans la petite région ou le domaine d'intérêt; le paramètre à estimer est donc le total de la variable x dans le domaine "a", ce que nous écrivons  $a_X$ .

Définissons  $a_h$  comme l'ensemble des unités appartenant à un domaine et situées dans la strate h. On peut donc écrire

$$a = \bigcup_{h=1}^L a_h \quad (2.1)$$

En pratique, l'intersection entre le domaine "a" et un certain nombre de strates définies par le plan de sondage ne sera pas nulle, et si nous désignons par  $\tilde{h}$  l'ensemble de ces strates, nous obtenons la formule

$$a = \bigcup a_h \cdot \tilde{h} \quad (2.2)$$

Nous examinerons un plan de sondage par grappes à plusieurs degrés, avec autopondération dans chaque strate, la strate h ayant un poids  $W_h$ . Pour un échantillon donné, nous pouvons calculer:

$$t_h = \text{total de la variable } x \text{ dans la strate } h$$

et



document sont adaptés à l'EPA canadienne qui est composée de domaines non prévus (type c) dont la taille est telle que, s'ils avaient été des domaines prévus (type a), la fiabilité des estimations ordinaires sans biais de l'enquête auraient été satisfaisantes au point qu'il n'ait pas fallu recourir à des techniques d'estimation pour des petites régions. En outre, les unités primaires d'échantillonnage de l'EPA sont petites (de 2000 à 5000 habitants) par rapport à la taille des domaines d'intérêt, contrairement à ce qui se fait aux États-Unis où la taille des unités primaires d'échantillonnage de la plupart des enquêtes à grande échelle est importante et comparable à la taille des petites régions pour lesquelles les estimations sont désirées.

Dans la présente étude, on évalue des estimateurs appliqués à la production d'estimations au niveau des divisions de recensement à partir de l'enquête sur la population active et à l'aide de données auxiliaires tirées des recensements de la population et du logement de 1971 et de 1976. En plus des estimateurs synthétiques, nous évaluons des estimateurs pour les domaines stratifiés a posteriori que Singh et Tessier (1976) ont déjà examinés, ainsi que des estimateurs composites formés de combinaisons linéaires des estimateurs synthétiques et des estimateurs pour les domaines stratifiés a posteriori, comme ceux examinés par Schabale (1979) et Schabale, Brock et Schnack (1977). Par ailleurs, nous proposons et évaluons un nouvel estimateur que nous appelons un estimateur dépendant de l'échantillon, qui a la même forme que l'estimateur composite, sauf que le poids de la composante synthétique est une fonction décroissante de la taille de l'échantillon compris dans le domaine, jusqu'à un point critique au-delà duquel l'estimateur dépend totalement de la composante relative au domaine stratifié a posteriori. L'efficacité des estimateurs pour les petites régions par rapport à celle des estimateurs directs (pour des domaines simples) a été mesurée pour les catégories des personnes occupées et des chômeurs au moyen d'une étude empirique (méthode de Monte-Carlo) dans laquelle on a simulé le plan de sondage de l'EPA en utilisant des données de recensement. On a examiné le cas où les données du plan d'échantillonnage et les renseignements auxiliaires sont tous à jour et celui où ils ne le sont pas. Nous avons évalué le biais des estimateurs synthétiques pour les catégories des personnes occupées et des chômeurs dans les circonscriptions électorales fédérales.

les cas qu'on peut trouver ici au pays, mentionnons les circonscriptions électorales fédérales, les divisions ou les subdivisions de recensement, les comtés et les régions de planification de la main-d'oeuvre.

Il convient de signaler que les types a) et c) correspondent à des domaines aréolaires.

Nous considérons que la répartition des domaines ci-dessus est importante puisque la forme aussi bien que l'efficacité d'un estimateur dépendent du type d'application en cause. Comme le font observer Purcell et Kish, la plupart des recherches sur les techniques d'estimation pour les petites régions effectuent aux États-Unis ou ailleurs ont porté surtout sur les domaines de type a) et b). Au Canada, par contre, ces deux types de domaines posent moins de problèmes qu'ailleurs en raison du plan de sondage proprement dit et de la taille des enquêtes nationales. On s'est donc penché davantage sur les données appartenant au type de domaine c), à l'exception peut-être des chiffres de population calculés à partir de données représentatives.

Les travaux d'application et d'évaluation de techniques d'estimation pour les petites régions utilisant des variables autres que la population ont débuté avec la publication d'estimations synthétiques du National Center for Health Statistics aux États-Unis, en 1968. Depuis lors, une série d'études (Gonzalez (1973), Gonzalez et Waksberg (1973), Schabie, Brock et Schnack (1977), Gonzalez et Hoza (1978) et d'autres) ont été menées à l'aide de données de l'enquête sur la population américaine (Current Population Survey) dans le but d'appliquer et d'évaluer un estimateur synthétique en particulier. Avec un estimateur synthétique de forme différente, Purcell et Linacre (1976) ont essayé de produire des estimations au niveau des divisions de recensement en Australie et, Ghangurde et Singh (1976, 1977, 1978) ont évalué des estimations synthétiques dans le cadre de l'enquête sur la population active du Canada (LEPA).

Comme le notent Purcell et Kish (1979), le rapport entre le caractère du plan d'échantillonnage et les domaines d'intérêt joue un rôle important dans le choix d'un estimateur. Les estimateurs dont il est question dans le présent

étudient le problème des estimations démographiques dans les petites régions en particulier, et ont trouvé plusieurs méthodes différentes fondées sur l'utilisation de données administratives ou d'autres sources.

Une analyse complète des techniques courantes d'estimation pour les petites régions (domaines d'étude) et des lacunes de ces méthodes a été faite par Purcell et Kish (1979). D'après les recherches effectuées jusqu'à présent, il semble bien ne pas y avoir de solution unique qui soit la meilleure pour régler le problème de l'estimation pour les petites régions. Le choix d'une méthode particulière pour ce genre d'estimation dépendra des besoins en données et de la richesse et de l'accessibilité des sources d'information, qui varient d'un pays à l'autre et, à l'intérieur d'un pays, selon le sujet étudié. La façon la plus convenable d'aborder l'analyse de données sur les petites régions serait donc de classer les types de petite région (ou de domaines d'étude), d'examiner les sources de données disponibles dans un contexte précis pour ensuite faire une étude détaillée des diverses techniques d'estimation applicables à une situation donnée. À ce propos, nous avons adopté la classification des domaines suggérée par Purcell et Kish (1979), tout en indiquant le type de domaine auquel les méthodes exposées dans le présent document se rapportent principalement.

a) Domaines prévus - pour lesquels des échantillons distincts ont été prévus, conçus et choisis. Dans le contexte canadien, de tels domaines pourraient être des régions économiques ou des secteurs de planification à l'intérieur d'une province, ou même une province entière.

b) Classes combinées - qui recourent le plan et les unités d'échantillonnage (et qu'on peut aussi désigner domaines caractéristiques) par exemple, les tranches d'âge selon le sexe, la profession, l'activité économique.

c) Domaines non prévus - qui n'ont pas été définis au moment de l'élaboration du plan de sondage et qui peuvent donc recouper les strates du plan ou les unités primaires d'échantillonnage (UPF) à l'intérieur des strates. Parmi



# ÉVALUATION DES TECHNIQUES D'ESTIMATION POUR LES PETITES RÉGIONS DANS L'ENQUÊTE SUR LA POPULATION ACTIVE AU CANADA<sup>1</sup>

J.D. Drew, M.P. Singh, G.H. Choudhry<sup>2</sup>

Il est parfois nécessaire d'avoir des estimations tirées d'enquêtes-échantillon pour des domaines d'étude dont les limites ne sont pas conformes à celles des strates prévues dans un plan de sondage. En prenant l'enquête sur la population active du Canada comme exemple d'une enquête fondée sur un plan d'échantillonnage en grappes, on évalue de façon empirique quelques techniques d'estimation pour les petites régions mises au point par divers auteurs. Parmi ces techniques, il y a les estimateurs synthétiques, pour domaines simples et pour domaines stratifiés a posteriori, ainsi que les estimateurs composites qui se forment par combinaison linéaire des estimateurs synthétiques et des estimateurs pour domaines stratifiés a posteriori. Un problème est un estimateur dépendant de l'échantillon qui affecte un poids à l'estimation obtenue pour les domaines stratifiés a posteriori en fonction de la taille de l'échantillon regroupé dans un domaine, et les résultats de cette méthode sont évalués.

## 1. INTRODUCTION

L'importance grandissante de la planification, de l'application et de la surveillance des programmes sociaux et fiscaux a suscité une demande accrue de données locales de bonne qualité de la part des administrations municipales, provinciales et fédérale, ainsi que du secteur privé. Les données requises peuvent varier de simples chiffres de population à des variables socio-économiques complexes comme l'emploi, le chômage, le revenu, le logement, les indices de pauvreté, l'état et les services de santé et d'autres encore. Jusqu'à récemment, pas beaucoup d'attention n'a été accordée à l'élaboration de bonnes techniques d'estimation de données sur les petites régions, sauf dans le cas des démographies statistiques qui, depuis un certain temps,

<sup>1</sup> Texte présenté à la réunion de l'American Statistical Association, à Cincinnati, en août 1982

<sup>2</sup> J.D. Drew, M.P. Singh et G.H. Choudhry, Division des méthodes de recensement et d'enquête-ménages, Statistique Canada

- [19] Platek, R.A. et Singh, M.P. (1981), Cost Benefit Analysis of Controls in Surveys", in Current Topics in Survey Sampling, New York, Academic Press, 1981.
- [20] Shannon, C.E. et Weaver, W. (1981) The Mathematical Theory of Communication, University of Illinois Press, Indiana, ILL.
- [21] Statistique Canada, (1981), Conception des questionnaires: Manuel d'atelier, 3e tirage, rapport non publié.
- [22] Sudman, S. (1980), Reducing Response Errors in Surveys. Statistician, 1980, Vol. 29, 237-273.
- [23] Warwick, D.P. et Lininger, C.A. (1975), The Sample Survey: Theory and Practice, New York, McGraw-Hill.



- [11] Jabine, T.B. (1981), Guidelines and Recommendations for Experimental and Pilot Survey Activities in Connection with the Inter-American Household Survey Program, Washington, D.C., Inter-American Statistical Institute report 7679a - 5/7/81 - 100.
- [12] Kahn, R.L. et Cannell, C.F. (1957), The Dynamic of Interviewing, Wiley and Sons.
- [13] Koch, G. (1973), An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means. Journal of the American Statistical Association, Vol. 68, N° 344, 906-913.
- [14] Marquis, K.H., Marquis, M.S. et Polich, J.M. (1981), Survey Response Errors for Sensitive Topics: The Problem is Noise Rather than Bias, présenté à la 141<sup>e</sup> Réunion annuelle de l'American Statistical Association, Detroit.
- [15] National Center for Health Statistics (1972), Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents, Vital and Health Statistics, Series 2, N° 50, Washington, D.C.: US Government Printing Office.
- [16] Neter, J. et Waksberg, J. (1965) Response Error in Collection of Expenditures Data by Household Interviews: An Experimental Study. Bureau of the Census Technical Paper N° 11, Washington, DC US Government Printing Office.
- [17] Payne, S.L. (1981), The Act of Asking Questions, Princeton University Press, Princeton, N.J.
- [18] Platek, R. (1980), Causes of Incomplete Data, Adjustments and Effects. Techniques d'enquête, vol. 6, n° 2, 93-132.

# BIBLIOGRAPHIE

- [1] Anderson, R., Kasper, J., Frankel, M.R. and Associates (1979), Total Survey Error, Jossey-Bass Publishers, San Francisco.
- [2] Ashraf, A., (1975), The Methodology of the Canadian Travel Survey. Techniques d'enquête, vol. 1, n° 2, 108-227.
- [3] Bailar, B. (1976), Some Sources of Errors and Their Effect on Census Statistics, Demography, vol. 13, n° 2, 273-286.
- [4] Bushery, J.M. (1981), "Recall Biases For Different Reference Periods in the National Crime Survey", présentée à la 141e Réunion annuelle de l'American Statistical Association, Detroit.
- [5] Carson, E.M., (1973), Questionnaire Design: Some Principles and Related Topics, Statistique Canada, document interne.
- [6] Chinnappa, B.N. et Wills, B. (1978), A Study of Refusal Rates to the Physical Measures Component of the Canada Health Survey, Techniques d'enquête, vol. 4, n° 1, 100-114.
- [7] Dillman, D.A., Mail and Telephone Survey: The Total Design Method, Wiley, New York.
- [8] Fellegi, I. (1979), Data, Statistics, Information - Some Issues of the Canadian Social Statistics Scene. Techniques d'enquête, vol 5, n° 2, 130-161
- [9] Groves, R.M. et Kahn, R.L. (1979), Surveys by Telephone, New York, Academic Press, 1979.
- [10] Hansen, M.H., Hurwitz, W.N. et Bershad, M.A. (1961), Measurement Errors in Censuses and Surveys, Bulletin of the International Statistical Institute, Vol. 38, Part II, 359-74.

peut poser différents problèmes et tendre des pièges, et il faut prévoir ces difficultés et en tenir compte dans la conception des questionnaires.

## 5. CONCLUSION

Ce document décrit le questionnaire comme pouvant représenter les besoins en renseignements des utilisateurs et comme un des principaux déterminants de la qualité des données d'enquêtes. À cause de ces deux fonctions, il y a interdépendance entre le questionnaire et tous les éléments nécessaires à la mise sur pied d'une enquête. Le plan de sondage global et, en particulier, le questionnaire, doivent produire des données exactes et les plus utiles possible pour les utilisateurs. Un questionnaire bien conçu répond à ces deux exigences; il faut poser la bonne question et il faut la poser de la bonne façon.

Il convient de souligner le fait qu'il y a parfois conflit entre les besoins de l'utilisateur et les exigences concernant l'exactitude des données. Le processus de mise au point d'un questionnaire comporte un certain nombre de compromis. Ainsi, il est possible qu'il faille formuler une question dans des termes plus simples que ne le ferait l'utilisateur pour s'assurer que le répondant peut donner une réponse. Par contre, il ne faut pas éviter de poser des questions complexes simplement parce que les réponses pourraient contenir des erreurs.

La conception d'un questionnaire n'est pas un simple travail de laboratoire. Bien qu'il existe certaines normes et que la recherche soit possible, l'art de construire un questionnaire s'apprend dans une bonne mesure par l'expérience pratique et par une série d'épreuves et d'échecs. C'est un métier qui s'apprend en discutant avec les utilisateurs des données, avec les interviewers et avec les enquêtes. Il s'agit clairement d'un processus interactif qui ne peut pas se dérouler isolément ou indépendamment des autres éléments de la préparation d'une enquête. Le questionnaire et ces éléments sont interdépendants et, en fait, il se trouve au centre de l'ensemble du plan de sondage.

L'inscription de codes directement sur le questionnaire en vue de la saisie des données, par exemple, a pour effet normalement de réduire considérablement les erreurs introduites pendant cette opération. Ainsi, les données peuvent être saisies directement à partir du questionnaire sans qu'il faille d'abord les transcrire sur une autre formule. Dans l'interview téléphonique automatique, la saisie des données est encore plus directe. En effet, le questionnaire est stocké dans un programme informatique qui contrôle le déroulement de l'interview au complet. Les questions apparaissent une à la fois sur l'écran d'un terminal et l'interviewer pose la question au répondant et entre la réponse directement dans l'ordinateur. Il est alors possible de vérifier les données immédiatement et de corriger les erreurs pendant que le répondant est encore au téléphone. Cette méthode permet aussi de diminuer le nombre de questions omises et celui des cas où on applique mal les instructions qui demandent de sauter certaines questions.

Il existe un lien étroit entre les erreurs de vérification et d'imputation et le contenu du questionnaire. Certains problèmes relatifs aux données qui manquent ou qui sont incohérentes peuvent souvent être attribués à des lacunes dans la conception du questionnaire. La capacité de reconstruire ou d'imputer des valeurs manquantes tient souvent aux autres variables incluses dans le questionnaire et aux mécanismes prévus pour éviter ce genre de difficulté. Par exemple, dans une enquête où on demande des renseignements sur plusieurs composantes détaillées du revenu, un bon nombre de cas où certaines données ne sont pas fournies ou sont erronées peuvent être récupérés si le questionnaire contient une question sur le revenu total.

Les erreurs de non-réponse, de réponse et de traitement sont quelques-unes des erreurs non dues à l'échantillonnage qui sont étroitement liées au questionnaire et aux autres éléments du plan de sondage global. Il est inévitable que le questionnaire soit une source d'erreurs d'observation, mais le questionnaire doit être conçu de manière à éliminer ces erreurs autant que possible. La mesure dans laquelle le questionnaire atteint cet objectif tient en grande partie aux connaissances que le responsable du plan d'enquête possède sur les diverses sources d'erreurs et à son aptitude à élaborer le questionnaire en fonction de tous les autres éléments de l'enquête. Chaque nouvelle enquête



influence sur tous les répondants qu'il rencontre, que ce soit par sa façon de poser les questions ou par sa manière d'interpréter et de noter les réponses, et ainsi de suite. L'importance de cette catégorie d'erreur par rapport à l'erreur totale de l'enquête est directement liée à la charge de travail confiée à un intervieweur. Dans les enquêtes téléphoniques, où le nombre de répondants attribué à chaque intervieweur peut être assez élevé, l'erreur correlative peut être beaucoup plus grave que dans le cas des interviews menées sur place (Groves et Kahn (1979)). Par contre, l'erreur de réponse correlative est plus sérieuse dans les interviews faites sur place que dans les enquêtes par la poste ou dans les enquêtes du type "autodéterminé". Ce problème constitue une des principales raisons pour lesquelles, depuis 1971, on utilise un système de livraison des questionnaires et de retour par la poste pour le recensement de la population et du logement. Ainsi, le choix de la méthode de collecte des données a un effet direct sur la forme du questionnaire.

On pourrait donner beaucoup d'autres exemples d'erreurs de réponse. Essentiellement, ces erreurs sont imputables au genre de questions posées, à la façon dont l'interviewer les pose, à la manière dont le répondant les interprète et y répond, et à la façon dont l'interviewer interprète et enregistre les réponses. L'interview est un processus de communication dynamique entre l'interviewer et le répondant. Le déroulement de l'interview détermine si les renseignements requis seront obtenus d'une façon efficace et avec exactitude. Au cours de l'interview, c'est surtout au questionnaire, par son contenu, par le libellé des questions, par les directives et par sa présentation, que doit revenir le rôle de contrôler la situation.

#### 4.3 Erreurs de traitement

Une fois que l'interview est terminée, le questionnaire devient surtout un document à dépouiller. Des erreurs peuvent survenir à toutes les étapes du traitement, notamment au codage, à la saisie des données, à la vérification, à l'imputation, à l'estimation et à la totalisation des résultats. La manière dont le questionnaire a été conçu a un effet marqué sur le nombre et les types d'erreurs produites à ce stade de l'enquête.



L'interview et de recueillir des renseignements supplémentaires corrélés aux variables d'intérêt. Chacune de ces solutions a des conséquences directes sur la conception du questionnaire.

Les questions qui reposent sur la capacité du répondant à se rappeler certains événements, comme un voyage ou un acte criminel, subi par le répondant, représentent une autre source d'erreurs de réponse. Des événements peuvent s'oublier, ou des événements qui se sont passés avant la période de référence peuvent être déclarés incorrectement. Busberry (1981), dans une expérience faite à l'aide de la "National Crime Survey" aux États-Unis, a constaté que les taux de victimes d'actes criminels au cours d'une période de référence de 3 mois étaient beaucoup plus élevés que ceux notés pendant une période de 6 mois, et que les taux observés dans cette dernière période dépassaient ceux calculés pour une période de référence de 12 mois. Bref, le biais dû aux problèmes de mémoire concernant les périodes de référence les plus longues était une source d'erreur plus importante que la variabilité d'échantillonnage. Le choix d'une période de référence appropriée pour les questions faisant appel à la mémoire a été examiné dans un certain nombre de domaines différents (Sudman (1980), National Centre for Health Statistics (1973)). La technique du rappel ordonné, où les répondants subissent une interview à la fin de la période de référence et où on se sert de dates marquantes (par exemple, Noël) ou de points de repère dans le calendrier afin de stimuler la mémoire des répondants, s'est avérée d'une certaine utilité pour diminuer les réponses incomplètes (Netter et Waksberg (1965), Ashraf (1975)). Toutefois, pour certains sujets, la seule façon possible de recueillir les renseignements nécessaires est de concevoir le questionnaire sous forme de carnet dans lequel le répondant consigne l'événement pendant qu'il se déroule ou peu de temps après. Ce genre de questionnaire est utilisé dans l'enquête sur les dépenses alimentaires et dans l'enquête sur la consommation de carburant de Statistique Canada.

Bien que les questions qui demandent un effort de mémoire ou qui portent sur un sujet délicat soient d'importantes sources d'erreurs de réponse, il existe un grand nombre d'autres causes. Par exemple, une des principales composantes de l'erreur de réponse est l'erreur due à l'interviewer, dite erreur de réponse corrélée. Chaque interviewer exerce, jusqu'à un certain point, une

Les erreurs de réponse sont la deuxième catégorie d'erreurs non dues à l'échantillonnage qui peuvent être imputables au questionnaire. Des erreurs de réponse peuvent survenir à n'importe quel moment pendant le cycle "question-réponse-enregistrement de la réponse", et il peut s'agir d'une erreur systématique (biais de réponse) ou aléatoire (variance de réponse).

Des questions sur des sujets délicats, comme la valeur et les sources du revenu, la consommation d'alcool et de tabac, les activités illégales ou la maladie mentale, ont tendance à produire de grandes erreurs de réponse. Par exemple, on a souvent l'impression que le répondant peut déformer sa réponse afin d'éviter la gêne ou pour donner l'impression qu'il se conforme aux normes sociales (Warwick et Linniger (1975)). Un grand nombre de méthodes d'élaboration de questionnaires ont été conçues pour tenter de réduire ce biais dû aux réponses axées sur ce qui semble socialement bon, comme le questionnaire anonyme, l'utilisation de questions projectives<sup>1</sup>, ou des techniques de réponse à des questions choisies au hasard, où le répondant lui-même choisit au hasard la ou les questions auxquelles il répondra parmi un choix de deux questions ou plus. Toutefois, dans une étude récente où l'on comparait les réponses obtenues au moyen de questionnaires et les renseignements de sources externes (par exemple, des dossiers administratifs ou des résultats de tests), Marquis et coll. (1981) ont découvert, à leur surprise, que pour la plupart des données qu'ils ont examinées le biais de réponse était presque négligeable, alors que la variance de réponse était assez grande. Cette observation, si d'autres études viennent l'appuyer, démontre qu'il est également important de mesurer et de diminuer la variance de réponse dans les enquêtes portant sur des sujets délicats. Il faudrait envisager comme solutions possibles de mener de nouvelles interviews, de vérifier la cohérence interne des réponses pendant

<sup>1</sup> Les deux questions suivantes donnent un exemple de questions projectives:

1. À votre avis, qu'est-ce que la plupart des gens pensent de l'usage de la marihuana?

2. Et qu'en pensez-vous?

La première question demande au répondant son impression de la norme sociale, tandis que la deuxième lui demande son avis personnel.

Afin de déterminer comment le questionnaire peut empêcher le refus, il est important de comprendre d'abord pourquoi les répondants acceptent ou refusent de participer. Un grand nombre de facteurs psychologiques motivent les gens à participer à une enquête, soit parce que le sujet les intéresse, parce qu'ils veulent se rendre utiles, parce qu'ils croient à l'utilité de l'enquête, parce qu'ils sentent un devoir de répondre, ou même parce qu'ils se croient importants. D'autres facteurs par contre, incitent les gens à refuser, par exemple la difficulté de comprendre les questions, la crainte des étrangers, la sensation de perdre son temps, la difficulté de se souvenir de certains éléments d'information, et les questions embarrassantes ou personnelles. Tous ces facteurs ont un effet sur la composition du questionnaire et influent sur la manière dont les sujets de l'enquête sont abordés, sur la formulation des questions, sur la présentation et la longueur du questionnaire, sur les garanties de confidentialité, et ainsi de suite. En même temps, il existe un rapport entre les facteurs psychologiques et le sujet de l'enquête, le type de population et la méthode de collecte des données. Ce sont là tous des éléments qui ont un effet sur le plan du questionnaire.

Il faut aussi prendre en considération les difficultés qui peuvent empêcher un répondant de répondre. Des questions qui manquent de réalisme parce qu'elles exigent des répondants des connaissances poussées ou des efforts de mémoire excessifs, l'utilisation d'un langage très compliqué ou très technique, ou des questions qui éprouvent la patience des répondants sont toutes des causes de non-réponse dont l'origine se trouve dans le questionnaire. Il faut dire, cependant, que la patience des répondants est souvent surprenante, même aux yeux des spécialistes endurcis. Chinnappa et Wills (1978) ont présenté une étude intéressante du problème de non-réponse à la partie de l'enquête Santé Canada qui portait sur les mesures physiques, étape où on demandait aux répondants de subir des tests de tension artérielle, des mesures du pli cutané, ainsi que de se soumettre à des exercices physiques et même des prises de sang.

Un exposé plus détaillé des causes et du traitement de la non-réponse a été présenté par Platek (1980).



l'échantillonnage, car elles comptent pour une part importante de l'erreur totale d'une enquête (voir, par exemple, Anderson et coll. (1979), Bailar (1976), Hansen, Hurwitz et Bershad (1961), Koch (1973), et Platek et Singh (1980)). Le traitement des erreurs non dues à l'échantillonnage est une partie intégrante et vitale du plan de sondage et il requiert des programmes spécifiques pour repérer, mesurer et éviter ces erreurs. De plus, chaque programme a ses propres coûts et ses propres avantages, facteurs dont il faut tenir compte dans l'élaboration des techniques de traitement (Platek et Singh (1980)).

Le questionnaire est à la fois une importante source d'erreurs d'observation et un élément important des programmes visant à éviter ou à mesurer ces erreurs. Le perfectionnement des méthodes de collecte des données a connu un certain retard par rapport à celui des plans d'échantillonnage et des techniques d'estimation. Dans bien des cas, l'amélioration des techniques d'échantillonnage ne vise que des erreurs d'environ une fraction d'un pour cent, tandis que des expériences sur la façon de formuler une question peuvent révéler des variations de vingt pour cent ou plus (Payne (1951)). La présente section décrit le lien entre le questionnaire et quelques-unes des sources les plus importantes d'erreurs non dues à l'échantillonnage et montre comment le questionnaire est censé réduire au minimum les erreurs.

#### 4.1 Erreurs dues à la non-réponse

La non-réponse est une des principales sources d'erreurs non dues à l'échantillonnage. Si les caractéristiques d'intérêt varient entre les répondants et les non-répondants, il est presque certain qu'une distorsion sera introduite dans les résultats. Il y a essentiellement deux types de non-réponse: le type "aucun contact" (par exemple, personne n'est à la maison, le répondant est temporairement absent, le temps est mauvais, etc.) et le type "refus". Ce dernier type peut comprendre une non-réponse totale ou le refus de répondre à certaines questions. Le questionnaire ne peut pas faire grand-chose pour éliminer la non-réponse de type "aucun contact", mais il peut faire beaucoup pour éviter le refus.

ce genre de modification comporte. Ainsi, il peut devenir impossible de faire des comparaisons dans le temps ou cela exige peut-être de recycler les interviewers ou de changer des programmes informatiques très coûteux.

Dans un bon nombre d'enquêtes permanentes, un peu comme dans l'enquête sur la population active au Canada, les mêmes répondants sont interviewés à plusieurs reprises. Le questionnaire doit prendre en considération l'ensemble du fardeau de réponse imposé à chaque répondant pendant qu'il demeure dans l'échantillon. Le questionnaire doit aussi pouvoir se prêter à diverses méthodes de collecte des données. Par exemple, la première interview de l'enquête sur la population active se fait toujours sur place, quoique dans bon nombre des régions urbaines la plupart des interviews subséquentes se font par téléphone. Les questionnaires des enquêtes permanentes doivent être élaborés dans une optique d'utilisation à long terme.

La conception du questionnaire est aussi liée au traitement des données et aux décisions d'ordre budgétaire. La structure des questions, par exemple si elles sont ouvertes ou fermées, a des incidences directes sur des opérations comme le codage, la saisie des données, la vérification et les totalisations. Si le questionnaire comporte beaucoup de questions ouvertes, cela exige plus de temps et de travail au codage, et rend la préparation et la mise à l'essai des programmes de vérification et de totalisation des données plus complexes et plus coûteuses.

Comme le questionnaire constitue une représentation opérationnelle des besoins des utilisateurs, il touche donc au plan de sondage global. Le plan de sondage est composé d'éléments complexes, parmi lesquels le questionnaire joue un rôle central. Le questionnaire ne détermine pas la forme des autres éléments, et ceux-ci ne déterminent pas la forme du questionnaire. Le processus de conception du questionnaire doit s'inspirer du processus d'élaboration du plan de sondage global et en faire partie intégrante.

#### 4. LE QUESTIONNAIRE ET LES ERREURS

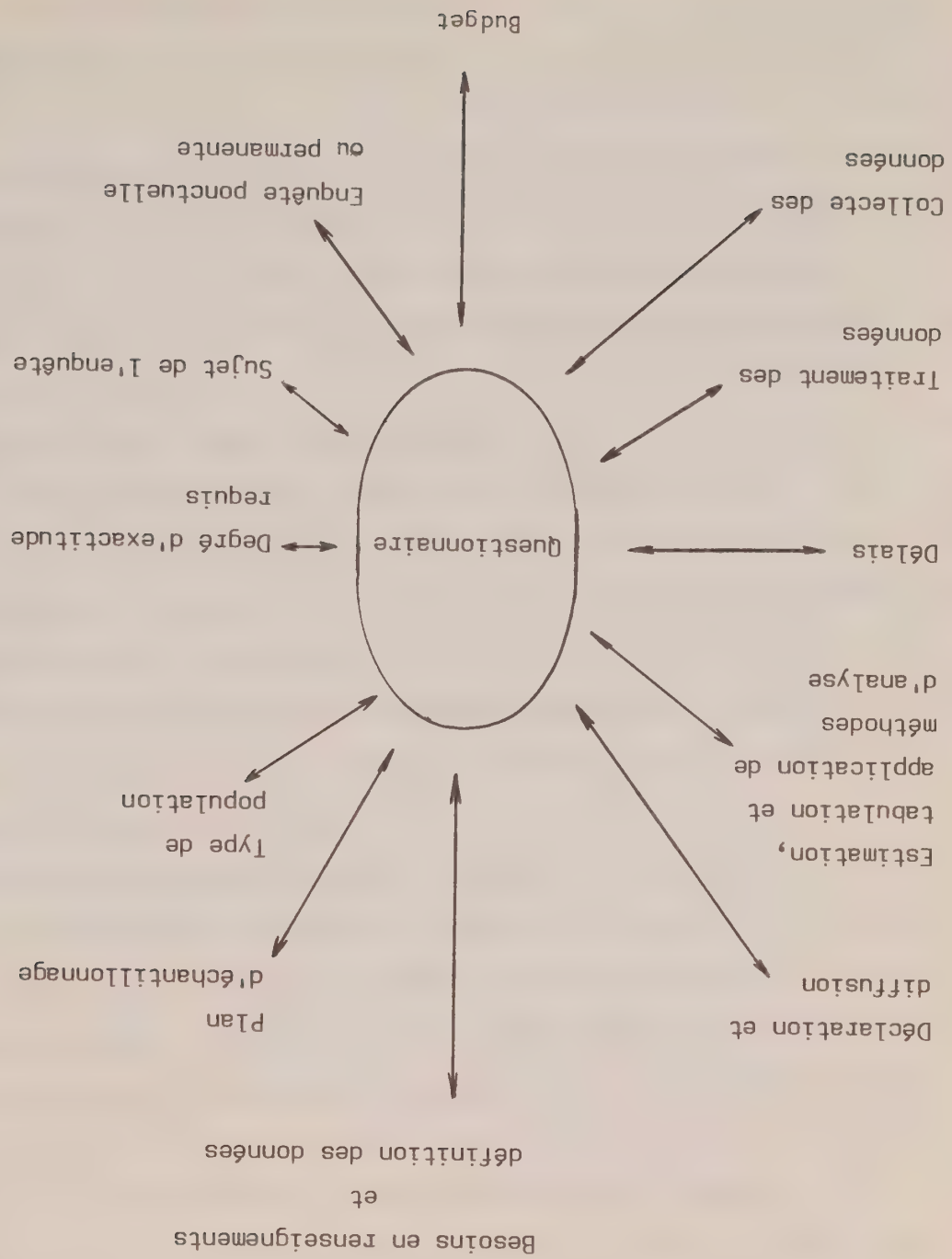
Toute enquête est entachée d'erreurs de diverses sources et, au cours des dernières années, on a accordé de plus en plus d'attention aux erreurs non dues à



La conception du questionnaire est liée de près à la méthode de collecte des données et au sujet de l'enquête. Chaque méthode de collecte, comme les entretiens sur place, les entretiens téléphoniques et les enquêtes par le poste, crée des conditions distinctes qui peuvent convenir plus ou moins à un sujet donné, et qui influent sur le genre de questions, ainsi que sur le contenu, le mode de présentation, la longueur du questionnaire, et ainsi de suite. Dans des entretiens sur place, par exemple, un intervieweur peut souvent recueillir certaines données, comme le type de logement et le sexe du répondant, par simple observation plutôt qu'en posant des questions. Par ailleurs, le questionnaire peut être conçu de façon à permettre à l'interviewer d'utiliser des cartes-questionnaire ou d'autres méthodes visuelles. Dans une interview sur place, la communication directe est un puissant facteur de motivation pour le répondant. L'interview sur place est souvent le seul choix qui se présente lorsque le questionnaire est complexe, long et exigeant. Dans les entretiens téléphoniques, une bonne partie de l'interaction sociale entre l'interviewer et le répondant est perdue, ce qui peut jouer sur la collaboration de ce dernier. L'efficacité du questionnaire, dans ce cas, repose exclusivement sur la communication verbale, et il se peut qu'il faille réduire la complexité du sujet de l'enquête. Cependant, dans certaines enquêtes sur des sujets délicats (par exemple, les enquêtes sur les victimes d'actes criminels), la distance entre l'interviewer et le répondant peut en fait favoriser la participation du répondant. Dans les enquêtes par le poste, le questionnaire remplace l'interviewer. Le document envoyé doit présenter l'enquête, motiver le répondant à collaborer et aider le répondant à fournir les renseignements. Le rôle de ce genre de questionnaire est particulièrement exigeant, et il faut en tenir compte au moment de sa mise au point.

Le caractère ponctuel ou permanent de l'enquête influe aussi sur la conception du questionnaire. Dans le cas d'une enquête permanente, il y a souvent plus de possibilités d'apprendre par expérience et d'améliorer le questionnaire avec le temps. Des recherches dans le domaine de la formulation des questions, l'application de programmes d'étude des erreurs de réponse ainsi que d'autres techniques d'évaluation et d'amélioration du questionnaire sont possibles seulement dans une enquête permanente. Toutefois, il faut examiner la possibilité d'améliorer un questionnaire en fonction des inconvénients que

Figure 1: Eléments qui influent sur le questionnaire



Ces trois problèmes s'appliquent directement à la construction des questionnaires, et ils sont tous liés de près. Dans le contexte des enquêtes statistiques, le questionnaire a une très grande influence sur la mesure dans laquelle les besoins des utilisateurs sont satisfaits.

### 3. LE QUESTIONNAIRE ET LES ÉLÉMENTS DU PLAN DE SONDAGE

Pour rendre les notions d'une enquête opérationnelles dans un document en particulier, le spécialiste est obligé de tenir compte non seulement de la formulation, de l'ordre et de la présentation de chaque question, mais également de presque tous les autres aspects de l'enquête. Dans la conception du questionnaire, il faut prendre en considération des facteurs comme le type de population visée, le plan d'échantillonnage et la taille de l'échantillon, le sujet de l'enquête, la méthode d'interview, les techniques informatiques qui seront appliquées aux données, ainsi que le budget et les délais.

La figure 1 illustre le lien entre le questionnaire et d'autres éléments du plan de sondage global. Les rapports entre ces éléments forment un réseau complexe. En effet, des modifications apportées à un aspect du plan de sondage entraînent souvent des changements dans plusieurs autres éléments. Il serait possible de mettre pratiquement n'importe quelle composante du plan d'enquête au centre de ce réseau mais, pour la présente étude, nous avons choisi le questionnaire comme point de départ.

Des éléments comme le type de population, le plan d'échantillonnage et le degré d'exactitude requis sont liés de près au plan du questionnaire. Par exemple, le caractère hétérogène de bon nombre de populations échantillonnées dans les enquêtes exige des classements recoupés de données. Cette exigence influe sur la taille de l'échantillon, le genre et le degré de stratification et la fiabilité de l'information, lesquels déterminent à leur tour le type de questions et le caractère plus ou moins détaillé des renseignements demandés. Les décisions prises à ce niveau ont des répercussions sur le coût et les délais de production des renseignements, sur le fardeau imposé aux répondants, et ainsi de suite.

rédigées en termes précis ni d'instructions claires à suivre, il est probable que les interviewers changeraient le sens ou la portée des questions et peut-être même les réponses. Le questionnaire permet d'assurer que le spécialiste mesure exactement ce qu'il veut mesurer chez tous les répondants. Il s'agit, en fait, d'un "programme" que l'interviewer et le répondant doivent suivre afin de produire le résultat voulu.

Cependant, le questionnaire doit être assez souple pour s'adapter aux répondants d'âge et de sexe différents, qui ne parlent pas la même langue et qui viennent de divers milieux sociaux. Il se peut qu'il faille utiliser des mots ou groupes de mots différents pour communiquer le bon message à tous les répondants. Le questionnaire doit aussi prévoir toutes les réponses possibles que les répondants peuvent donner. Ce principe est surtout vrai dans les phases exploratoires d'une recherche, quand la collecte d'un ensemble de données non structurées peut se révéler la meilleure méthode.

Il faut admettre que le questionnaire est un instrument de mesure complexe et souvent inexact. Les enquêtes sont des êtres humains, et le processus par lequel les mesures sont recueillies est fondé sur le langage. En plus de servir d'instrument de mesure, le questionnaire représente une forme de communication entre le spécialiste, l'interviewer et le répondant. Il transmet une demande de renseignements au répondant, et il envoie les réponses au spécialiste sous une forme que ce dernier peut traiter. Warren Weaver, dans son livre The Mathematical Theory of Communication (1949), définit trois problèmes dont il faut tenir compte dans la conception de tout système de communication :

A. Avec quel degré d'exactitude les symboles de communication peuvent-ils être transmis? (le problème technique)

B. Dans quelle mesure les symboles transmis communiquent-ils le sens voulu? (le problème sémantique)

C. Avec quel degré d'efficacité le message reçu influence-t-il sur le comportement, de la façon escomptée? (le problème d'efficacité)



Dans les enquêtes très structurées, le questionnaire offre un moyen de normaliser et de guider la collecte des données. Dans les sondages statistiques, au contraire des autres formes d'enquête, le spécialiste ne peut habituellement pas effectuer la collecte de données lui-même, mais il doit compter sur des interviewers embauchés à cette fin. S'il n'y avait pas de questions

Souvent, le questionnaire sert aussi de support sur lequel les mesures sont enregistrées. Un tel système est avantageux surtout pour les interviewers et les répondants, car il est commode de pouvoir inscrire une réponse juste après la question. Théoriquement, il n'y a toutefois aucune raison de ne pas écrire les réponses sur une autre formule que celle du questionnaire.

Une fois ces besoins définis sous forme de notions précises, le questionnaire devient l'instrument par lequel ces notions sont mesurées. Par des questions précises et des instructions appropriées, l'utilisateur indique explicitement comment les notions de l'enquête doivent être mesurées sur le plan opérationnel. Il faut parfois prévoir plusieurs questions pour mesurer certaines notions complexes. Par exemple, l'enquête sur la population active au Canada comprend une dizaine de questions servant à évaluer la notion de chômeur.

Il est important de souligner que le choix des termes utilisés dans les questions ne fait pas partie de l'élaboration des notions et définitions d'une enquête. Quand l'utilisateur détermine ses besoins en renseignements, il doit d'abord décider ce qu'il faut mesurer et non comment le mesurer. L'utilisateur doit choisir les notions qui conviennent le mieux à ses besoins, en examinant, par exemple, quelles notions semblent les plus appropriées, compte tenu de l'utilisation finale des données, et en déterminant la compatibilité des notions qu'il veut utiliser avec celles d'autres sources de renseignements.

notions utilisées dans une enquête décrivent à la fois ce qui doit être mesuré et les unités pour lesquelles il faut obtenir ces mesures. L'utilisateur peut définir "la situation du logement" en fonction du nombre de pièces, de la présence de certains éléments comme la plomberie ou l'électricité, ou de l'état du logement. Il peut définir "les pauvres" d'après le niveau du revenu ou selon la valeur de l'actif et des dettes.



## 2. LES BESOINS EN RENSEIGNEMENTS ET LE RÔLE DU QUESTIONNAIRE

La définition la plus simple d'un questionnaire est qu'il s'agit d'un groupe ou d'une suite de questions visant à recueillir des renseignements sur un sujet auprès d'un répondant. Comme divers types de questions sont possibles, le questionnaire peut être composé d'une liste de sujets non définis ou, à l'autre extrême, de questions très structurées ne permettant pas de répondre autrement que par le choix des réponses fournies.

Le questionnaire joue un rôle fondamental dans ce processus complexe (l'inter-vieu) où des renseignements sont transmis de ceux qui les possèdent (les répondants) à ceux qui en ont besoin (les utilisateurs). Le questionnaire est le moyen par lequel les besoins des utilisateurs en matière de renseignements sont exprimés de telle sorte que le répondant accepte de fournir l'information requise. Pour que cet échange de renseignements soit efficace, le questionnaire doit satisfaire les exigences des utilisateurs ainsi que des répondants. Un énoncé clair des besoins en renseignements, dont l'utilisateur peut n'avoir qu'une vague idée au début, n'est pas le résultat d'une seule opération. En réalité, le plan du questionnaire subit toutes sortes de transformations au cours du processus général d'élaboration d'une enquête.

Par exemple, un utilisateur désire obtenir des renseignements sur "la situation du logement chez les pauvres". Il traduit ce besoin en objectifs d'enquête en posant des questions comme:

a) Quel est le problème que nous tentons de résoudre?

b) De quels éléments d'information avons-nous besoin?

c) Comment les données seront-elles utilisées?

d) Jusqu'à quel point les renseignements doivent-ils être exacts et actuels?

Lorsqu'il répond à ces questions, l'utilisateur se met à penser de façon quantitative, et il formule ses besoins sous forme de notions précises. Les

## IMPORTANCE DU QUESTIONNAIRE DANS LE PLAN DE SONDAGE

R. Platek et D. Royce<sup>1</sup>

L'enquête statistique moderne est un outil très utile pour satisfaire la demande toujours croissante de données actuelles et exactes, et le questionnaire constitue un des éléments importants de l'enquête. Le présent document décrit le rôle du questionnaire par rapport aux besoins des utilisateurs, le lien entre le questionnaire et les autres éléments du plan de sondage, et l'effet du questionnaire sur la qualité des données recueillies. On souligne également l'importance de considérer le questionnaire comme une partie intégrante de l'ensemble du plan d'enquête.

### 1. INTRODUCTION

L'intérêt croissant manifesté pour des renseignements très variés qui soient significatifs et diffusés rapidement, à partir de nombreuses sources, exige une méthode structurée pour tout le processus de la collecte des données. Au cours des quarante dernières années, on a vu l'enquête statistique s'affirmer comme un instrument très utile pour répondre à cette demande.

Le questionnaire est un élément important de l'enquête statistique. Dans ce document, nous décrivons la capacité du questionnaire à satisfaire les besoins en renseignements, le lien entre le questionnaire et les autres éléments du plan de sondage, et l'effet du questionnaire sur la qualité des données recueillies. Bien que cet exposé se rapporte surtout aux enquêtes-ménages faites au moyen d'interviews sur place, bon nombre des observations s'appliquent également aux questionnaires et aux enquêtes de toutes sortes.

<sup>1</sup> R. Platek et D. Royce,  
Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada



Préparé par les méthodologistes de Statistique Canada

Comité de rédaction:	
R. Platek	- Président
M.P. Singh	- Rédacteur en chef
P.F. Timmons	
J.H. Gough	- Rédaction adjoint

Politique de la rédaction:

La revue "Techniques d'enquête" veut donner aux personnes qui s'intéressent aux aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquêtes: les problèmes de conception causés par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de documents pour publication

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages Statistique Canada, 4<sup>e</sup> étage, Édifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à inter-

ligne et demi.





TABLE DES MATIÈRES

Importance du questionnaire dans le plan de sondage	R. PLATEK, D. ROYCE .....	1
---	---------------------------	---

Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada	J.D. DREW, M.P. SINGH, G.H. CHOUDHRY .....	19
---	--	----

Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active du Canada	ELIZABETH CLAYTON PAUL, MURRAY LAWES .....	53
---	--	----

Le biais de renouvellement de l'échantillon dans les estimations de l'EPA	P.D. GHANGURDE .....	94
---	----------------------	----

Information du calcul d'estimation pour les enquêtes complexes	M.A. HIDIROGLOU .....	112
--	-----------------------	-----



# TECHNIQUES D'ENQUÊTE

1982

volume 8

numéros 1 & 2

Préparé par les  
méthodologistes de  
Statistique Canada

Canada



12-001

Government  
Publications



Statistics Canada Statistique Canada

---

# **SURVEY METHODOLOGY**

---

## **1983**

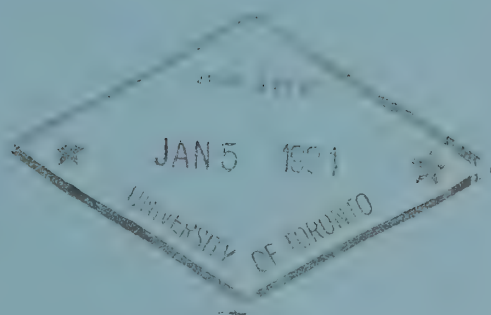
---

### **Volume 9**

---

### **Number 1**

---



---

A Journal produced by  
Statistics Canada

---

Canada





# SURVEY METHODOLOGY

1983

Vol. 9

No. 1

A Journal produced by Statistics Canada.

## CONTENTS

The Redesign of a Survey to Measure Commodity Origin and Destination Movements by the For-Hire Trucking Industry in Canada ROBERT LUSSIER and STEVEN MOZES.....	1
The Methodology of the Canadian Air Scheduled International Passenger Origin and Destination Estimation System GREG HUNTER and LISA DIPIÉTRO.....	27
Some Aspects of Quality of Cancer Mortality and Incidence Statistics D. BINDER and A. MALHOTRA.....	50
Estimating Monthly Gross Flows in Labour Force Participation STEPHEN E. FIENBERG and ELIZABETH A. STASNY.....	77
Redesign of the Niagara Tender Fruit Objective Yield Survey J. KOVAR.....	103
A Timely and Accurate Potato Acreage Estimate from Landsat; Results of a Demonstration R.A. RYERSON, J.-L. TAMBAY, R.J.BROWN, L.A. MURRAY and B. MCLAUGHLIN.....	119
Sampling on Two Occasions with PPSWOR G.H. CHOUDHRY and JACK E. GRAHAM.....	139

8-3200-501  
Reference No.  
Z - 079

ISSN: 0714-0045



# SURVEY METHODOLOGY

1983

Vol. 9

No. 1

A Journal produced by Statistics Canada.

---

Editorial Board:	G.J.C. Hole	
	C. Patrick	
	R. Platek	- Chairman
	M.P. Singh	- Editor
	P.F. Timmons	
	H. Lee	- Assistant Editor

---

## Editorial Policy:

---

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed; however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department.

## Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.



## THE REDESIGN OF A SURVEY TO MEASURE COMMODITY ORIGIN AND DESTINATION MOVEMENTS BY THE FOR-HIRE TRUCKING INDUSTRY IN CANADA<sup>1</sup>

Robert Lussier and Steven Mozes<sup>2</sup>

This paper firstly provides an overview of the For-hire Trucking Survey background and of the steps that were involved in the revision that led to its re-design. It secondly describes the general direction of the methodology of the re-designed survey which is being implemented for reference year 1981.

### 1. INTRODUCTION

The For-hire Trucking Survey was initiated by Statistics Canada in 1971 to measure commodity origin and destination movements by the for-hire trucking industry in Canada. For the purpose of the survey, this industry was defined as the sum of trucking establishments engaged in transportation of freight for compensation. The survey was a probability sample survey of shipments recorded on the shipping documents retained by Canada's for-hire trucking firms. Since 1971, the demand for more reliable and more detailed information has been increasing steadily. This increased demand can be attributed to a number of factors such as the dramatic growth of trucking activity since the early fifties; the increased sophistication of users of transportation statistics; the growing interest in the subject of economic regulation versus deregulation, and finally the increased market share of trucking, at the expense of other modes of transport, within the overall freight transportation market. In the late 1970's, Statistics Canada in cooperation with major users embarked on a complete revision of the survey.

---

<sup>1</sup> This is a revised version of an invited paper presented at the Joint Statistical Meetings of the American Statistical Association, the Biometric Society, ENAR and WNAR, and the Institute of Mathematical Statistics, Cincinnati, August 16-19, 1982.

<sup>2</sup> Robert Lussier, Business Survey Methods Division and Steven Mozes, Transportation and Communications Division, both of Statistics Canada.



It is the intention of this paper to serve two purposes. Firstly, it provides an overview of the background of the survey and a description of the steps in the revision process, and secondly, it describes the methodology of the redesigned survey which is being implemented for the reference year 1981. It should be noted that the details of the methodology of some phases have not yet been finalized; this paper will, however, include descriptions of the general direction of the incomplete phases.

## 2. BACKGROUND

### 2.1 Brief overview of the Canadian for-hire trucking industry

The Canadian for-hire trucking industry is characterized by a very large number of small operators and by a high degree of heterogeneity as manifested by the variety of commodities carried, size of operators, and area of operations.

Small carriers defined as carriers earning less than \$100,000 represent 88% of the industry, measured in terms of numbers; however, they represent only 20% of the industry when measured in terms of operating revenues. The existence of this large number of operators, their volatility and their relative insignificance in terms of revenues lead to the decision to exclude them from the survey population.

Trucking firms are involved in the transportation of widely differing commodities, requiring different kind of equipment and operating practices. The various carrier types (e.g. general freight, bulk petroleum, household goods movers, etc.) differ from each other not only in terms of the commodities they carry but also in terms of shipment size.

Heterogeneity can also be illustrated by describing the area of operation. Some trucking firms operate locally only, others intraprovincially, and some of the larger carriers, in each province as well as internationally.

The combination of these factors have implications on the survey design especially on stratification.

## 2.2 History of Canadian truck origin and destination surveys

### (a) The Motor Transport Traffic Survey (1957-1963)

The first attempt to measure truck traffic in Canada was made in 1957 with the introduction of the Motor Transport Traffic Survey (MTTS). This survey was a sample survey of motor vehicles engaged in freight transportation. The survey frame was a list of registered motor vehicles. It originated from the motor vehicle registration files maintained by the provincial or territorial governments. This frame was stratified by type of operation and gross vehicle weight.

The sample size was approximately 10% of all registered vehicles. The sample was selected in four quarterly segments with approximately one fourth of the total sample selected each quarter. Each quarterly sample was spread over three survey weeks with one third of the sample being used for a seven day period per month.

As the survey was conducted on a vehicle basis, no information was requested regarding the detailed origin and destination of the commodities carried. It was a truck origin and destination survey; commodities had secondary importance. Data relating to the vehicle such as the description of the vehicle, miles travelled, fuel consumed and the operating cost associated with the vehicle were also collected.

The survey was in operation from 1957 to 1963 inclusive. It was discontinued in 1964 because of changes in vehicle licensing systems, structural changes in the industry, and most importantly because of a very serious deterioration in response rates.

(b) The For-hire Trucking Survey (1969-1979)

Initial work on a survey to measure domestic intercity origin and destination traffic movements of goods by the total Canadian for-hire trucking industry began in 1969. At that time, a study of various methods of collecting commodity origin and destination statistics was carried out. The study results showed that a sample survey of the carriers' administrative records, namely their shipping documents, was a viable approach to the collection of the required data.

In 1970, a pilot survey was conducted to assess the effectiveness of the survey approach. The pilot survey involved the examination of the shipping documents of 187 for-hire trucking firms throughout the country.

The favourable response to the pilot survey and the availability of origin, destination, commodity, weight and revenue information on the shipping documents indicated that the survey approach was feasible.

For-hire trucking surveys were therefore conducted for reference years 1970 and 1971 with the above-mentioned objective. For reference year 1972, the objective was modified to restrict the survey to Canadian domiciled for-hire carriers earning \$100,000 or more annually from inter-city trucking. For reference year 1973, an updated and better-defined frame of regulated motor carriers was used and a more effective sampling procedure was developed. Since reference year 1973, the survey has been conducted and the results published on an annual basis by the Transportation and Communications Division of Statistics Canada [1] [5].

### 3. REVISION OF THE FOR-HIRE TRUCKING SURVEY

The revision process consisted of two main phases. Firstly, a critical review of the existing survey was initiated. Secondly, on the basis of the recommendations made during the review process, a complete survey redesign was

undertaken.

### 3.1 Survey review

#### (a) Reasons for the review

In early 1978, Statistics Canada initiated a review of the For-hire Trucking Survey for the following reasons. First, it has been the policy of the Transportation and Communications Division of Statistics Canada to conduct a periodic review of each of the ongoing surveys. The For-hire Trucking Survey had not been reviewed since 1973. Secondly, the current and anticipated future needs of users for increased details for commodity origin and destination statistics could not be satisfied within the constraints of the survey. Thirdly, experience gained during the undertaking of the For-hire Trucking Surveys and other related surveys provided additional information upon which the frame, the stratification variables and the imputation techniques could be improved. Fourthly, some developments in the trucking industry, such as the availability of origin and destination information in machine readable format lead to the belief that computer tapes could be utilized to increase the data base and reduce reporting burden at the same time.

In addition, increased sophistication of users required the development of improved data dissemination procedures while changes in data processing technology had made the present data processing system not only obsolete from a technical point of view but cost inefficient as well.

#### (b) Phases of the review

The survey review was originally organized into two parts, namely Phase I and Phase II.

The objective of Phase I was to outline recommendations which concentrated on improvements to the survey within its existing framework using only limited additional resources. The recommendations had to focus and indeed did focus



on a redefinition of the survey population, improvements in stratification variables and an increase in the sample size of shipping documents. The recommendations were presented in a report [2].

The objectives of Phase II were to assess the Phase I recommendations from a user point of view, to present various cost and implementation alternatives for the accepted recommendations and to complete further survey analyses. Phase II reformulated some of the Phase I recommendations and added additional recommendations which aimed at a smaller population of firms better stratified into more homogeneous groups. In addition to recommending the implementation of these recommendations, four alternatives for increasing the sample size were considered namely, the status quo; an increase of 50% to the sample of shipping documents; an increase of 100% to the sample of shipping documents; and finally an increase of 25% to the sample of shipping documents together with the processing of available carrier data tapes for presumably 40 or so firms. Based on an assessment of the advantages and costs of each of these alternatives, the latter one was approved in principle because it offered the potential for substantial sample size increases with a minimum of cost and data collection burden. The recommendations and the supporting details were tabled in a report [3].

A preliminary assessment of the impact of the recommendations revealed that further work was necessary especially to determine the full costs for the use of carrier waybill tapes. Therefore a Phase III was added to the survey review process. In general, its terms of reference were to conduct the investigations required to formulate and recommend general specifications for a revised survey. The investigations had to follow the recommendations of the Phase II review.

In June 1980, Phase III proposed that the survey be redesigned to accept four types of input namely, tapes from selected respondents; transcriptions from samples of shipping documents drawn from each Document Storage Location Point (D.S.L.P.) having over 1.5 million intercity domestic revenue annually; transcriptions from samples of shipping documents drawn from a sample of D.S.L.P.'s



having between \$350,000 and \$1.5 million intercity domestic revenue annually; and macro information from D.S.L.P.'s with annual intercity domestic revenue between \$100,000 and \$350,000. The decision to collect macro information from the smaller carriers was based on that fact that these firms do not keep the documentation needed for sampling purposes.

### 3.2 Survey redesign

#### (a) Objective of the redesign

After the completion of Phase III of the survey review process, it was decided to carry out a complete redesign of all aspects of the survey. The objective of the redesign was to provide more reliable and more detailed commodity origin and destination statistics relating to the Canadian for-hire trucking industry. It was expected that both the reliability and the amount of regional and commodity detail available could be increased when compared with the "old survey".

#### (b) Constraints on the redesign

The main constraints imposed on the redesign were: that the survey population exclude some types of for-hire trucking firms namely, own account household goods carriers and oil field carriers; that the stratification be improved to be more in line with the economic structure of the industry; that three types of input be accepted, namely, tapes from selected respondents, transcriptions from samples of shipping documents drawn from D.S.L.P.'s of a sample of firms having more than \$350,000 intercity domestic revenue annually and macro-information from a sample of firms having annual intercity domestic revenue between \$100,000 and \$350,000; and finally that the redesigned survey be implemented for the reference year 1981, data collection starting in the spring of 1982.

#### 4. POPULATION AND FRAME

The population of the survey covers all shipments made during the reference year by those trucking firms which are defined as in scope for the survey. A shipment is defined as a quantity of merchandise transported by the carrier from one person or organization to another person or organization. The in-scope firms include those which earn more than \$100,000 annually from intercity freight transportation, whose main activity is trucking and who are Canadian domiciled. Excluded from this population are the shipments of certain types of specialized carriers such as the oilfield carriers and own account household good movers.

However, this ideal population is not accessible. As a substitute, firms are used as natural clusters of shipments for the first-stage sampling units of the design.

For this reason, the frame consists of a list of all firms which have domestic intercity revenue over \$100,000. Firms may further be segregated into D.S.L.P's. This is the case for those firms whose shipping documents are not stored at a central location. The frame is derived from an annual census survey of for-hire trucking conducted by Statistics Canada, the Motor Carriers-Freight and Household Goods Movers Survey<sup>3</sup>.

#### 5. ULTIMATE SAMPLING UNIT

The survey accepts three different inputs, namely tapes from selected carriers, transcribed information from sampled shipping documents and finally macro information from carriers earning between \$100,000 and \$350,000 annually.

---

<sup>3</sup> The Motor Carriers-Freight and Household Goods Movers Survey of Statistics Canada is an annual census survey of trucking establishments. Its objective is to obtain establishment-oriented input-output data such as revenues, expenses, balance sheet information and equipment operated.

The tapes contain information relating to individual shipments, the characteristics of which are the same as those which are recorded on shipping documents. Therefore, the ultimate sampling unit for those firms which either provide tapes or whose shipping documents are sampled is the shipment. For the carriers in the \$100,000 to \$350,000 range, macro information is obtained as these firms do not usually keep the necessary documentation relating to shipments. For these carriers, the ultimate sampling unit is the firm.

## 6. INFORMATION COLLECTED

The principal characteristics needed from each shipment sampled from carriers earning more than 350,000 intercity domestic revenue annually are the true origin and the final destination; the description of the commodity(ies) carried; the weight and the unit of weight; the transportation revenue earned and the interlined shipment information. Interlining occurs when a consignment is moved by a carrier to an intermediate point and then moved by another carrier to another point. The interlined shipment information is used to eliminate duplications.

The secondary characteristics needed are the date of shipment; the quantity of commodity and the unit of measurement (e.g. 5 board feet, 20 gallons, 15 sacks); some information regarding the shipment weight transcribed (e.g. minimum weight, convenient weight used for calculating revenue); the rate charged and the rate condition codes (e.g. a code indicating where rate is minimum, per 100 lb., per hour) and the revenue condition codes. (e.g. a code indicating where exact transportation revenue is not available, where the shipment is out-of-scope).

The macro information collected from the smaller carriers describe the average or typical shipments in terms of originating province, destination province, commodity, average revenue, average weight and number of shipments.

## 7. ADMINISTRATIVE RESTRICTIONS

The amount of resources available for data collection and processing and the goal to reduce the burden imposed on the respondent put a limit on the number of firms selected and on the number of shipments selected and transcribed.

### 7.1 Maximum number of firms in the sample

The population of the 1981 For-hire Trucking Survey<sup>4</sup> consists of 2,711 firms of which 1,288 earn more than \$350,000 annually while 1,423 earn between \$100,000 and \$350,000 annually.

As data collection is very expensive due to the very high cost associated with travelling to remote areas, efforts are being made to limit the number of D.S.L.P.'s selected in the sample from those carriers earning over \$350,000 annually. The limit is set at 875 D.S.L.P.'s per year, which has been the historical number during the last ten years of the old survey.

### 7.2 Maximum total number of transcriptions

The second administrative restriction relates to the total number of transcriptions. The present budget allocation allows a maximum of 418,000 transcriptions. This number may vary from year to year depending on negotiations between Statistics Canada and users who are also cofinancers of the survey.

### 7.3 Maximum number of transcriptions per firm

There is also an administrative restriction which relates to the maximum number of transcriptions per firm. There is an implicit limit imposed on the number of days the data collection team can spend at any particular location, so that the respondents are not burdened by the presence of the Statistics Canada regional operations personnel.

---

<sup>4</sup> 1981 For-hire Trucking Survey means the survey conducted in 1982 for reference year 1981.



## 8. STRATIFICATION AND SAMPLE ALLOCATION

Using the results of the previous year's Motor Carriers-Freight and Household Goods Movers Survey, the firms are stratified according to their in-scope transportation revenue, type of operation and area of operation. These variables were chosen because they characterize the heterogeneity of the industry. The in-scope transportation revenue indicates if the firm is a Class 1, a Class 2 or a Class 3 firm i.e. if the firm earned \$2.7 million or more, between \$350,000 and \$2.7 million or between \$100,000 and \$349,999 dollars of revenue respectively from Canadian intercity non-armoured and non-household goods freight transport. The type of operation characterizes the firms as specializing in general freight small shipments, general freight large shipments, automobiles, liquid petroleum, dump trucking, forest products, building materials, dry bulk and/or refrigerated liquids, heavy machinery, refrigerated solids, explosive and/or other dangerous goods, agricultural products, animals and van lines. The general freight small shipment carriers are general freight carriers for which the average revenue per shipment is less than \$85.00; the general freight large shipment carriers are the rest of the general freight carriers. The area of operation indicates the specific Canadian province, Yukon or Northwest Territories, or that combination in which the firm operates. For example, an area of operation could be New Brunswick, meaning that the firm operates in New Brunswick only. Another example would be Atlantic which means that the firm operates in 2 or more of Newfoundland, Prince Edward Island, Nova Scotia and New Brunswick but nowhere else in Canada. There are 20 of these areas of operation.

The dollar cut-offs used in the stratification by revenue and by type (i.e. \$85, \$350,000 and \$2.7 million) are flexible and may vary in the years to come depending on the changes occurring in the population.

The above stratification creates 840 strata of which 355 were non-empty in the 1981 For-hire Trucking Survey.

Once the frame is stratified, subject matter officers may identify take-all



firms i.e. firms that they want to be included in the sample with probability one. Next, a methodologist determines the number of firms to be selected among the non take-all firms in the stratum. To do so, he goes through several steps from which the take-all firms are excluded.

First, a computer program calculates the initial number of firms to be selected in each stratum to meet a target coefficient of variation of the estimate of in-scope revenue in the stratum. This target coefficient of variation is the coefficient that one would like to obtain if the estimate were calculated using the reported total revenue from a sample of firms selected using simple random sampling from a population of firms for which the distribution of the in-scope revenue is the same as the distribution of the in-scope revenue of the previous year's Motor Carriers-Freight and Household Goods Movers Survey. The formula is:

$$n_{1h} = \frac{N_h^2 S_h^2}{N_h S_h^2 + Y_h^2 (C.V._h)^2}$$

where  $n_{1h}$  : initial number of firms to be selected among the non take-all firms in stratum h;

$N_h$  : number of non take-all firms in stratum h;

$Y_h$  : total in-scope revenue of the non take-all firms in stratum h;

$S_h^2$  : variance of the in-scope revenue of the non take-all firms in stratum h; and

$C.V._h$  : target coefficient of variation in stratum h (the value used is the same for all strata of a given class but may vary from class to class).

Secondly, the initial sample sizes are revised to ensure that a minimum number of firms is selected from each stratum i.e.

$$2n_h = \min \{ \max (m, 1n_h), N_h \}$$

where  $2n_h$  : revised initial number of firms to be selected among the non take-all firms in stratum h; and

m : minimum number of firms to be selected in stratum h if possible<sup>5</sup>.

Then the revised initial sample sizes are summed over the strata to get a total revised initial sample size.

Next, the sample sizes are again reviewed to ensure that the sample size in a given stratum is greater or equal to the sample size that one would have obtained if he had distributed the total revised initial sample size of a class across the strata of the class proportionally to the square root of the number of firms in each stratum i.e.

$$3n_h = \max \left\{ \frac{\sqrt{N_h}}{\sum_h \sqrt{N_h}} \sum_h 2n_h, 2n_h \right\}$$

where the summation is done over all strata of the same class than stratum h.

Finally, the survey manager may subjectively adjust the sample sizes to  $4n_h$ .

The above sample allocation method has been retained because it is an algorithm which has given satisfactory results during the testing phase as well as has made use of the only variable that was available for all firms namely the in-scope revenue of the firms. Nevertheless, it should be realised that the in-scope revenues of the firms are not collected directly in the For-hire Trucking Survey but revenues from a sample of shipments are collected. Therefore the above method of sample allocation ignores completely the second stage of sampling.

---

<sup>5</sup> For reference year 1981, this minimum was set to 3 for all strata.

## 9. FIRST STAGE SAMPLE DESIGN

The first stage consists of selecting in each stratum a number of firms corresponding to the number of firms  $4n_h$ , determined at the sample allocation stage.

All firms earning \$2.7 million or more of in-scope transportation revenue were made take-all i.e. were selected with probability one in the 1981 For-hire Trucking Survey. The reason for this approach is that these firms are known to be heterogeneous with respect to the principal statistics to be estimated and are known to be contributing a large proportion of the revenue figures to be estimated.

The sample of firms is finally converted to a sample of D.S.L.P.'s by including in the latter sample all D.S.L.P.'s of the selected firms.

## 10. SECOND STAGE SAMPLE DESIGN

The second stage of the sample design for D.S.L.P.'s of Class 1 and Class 2 firms consists of selecting a systematic sample of shipments from the files of each selected D.S.L.P. This selection is done by Statistics Canada Regional Operations Division interviewers at the D.S.L.P. The sampling intervals used are different depending on the number of shipments carried by the firms. They are generally obtained from a table provided to the interviewers. This table gives various file size ranges with their corresponding sampling interval. However, the sampling intervals may be pre-determined for any given firm by Statistics Canada Head Office staff. This is especially the case of multi-D.S.L.P. firms because the interviewer at a given D.S.L.P. may not know how many shipments were carried by the firm as a whole. This is also the case for firms having special characteristics, such as firms carrying dangerous goods, and others for which the survey manager may want a larger data base. In subsequent years, this may also be the case for firms contributing to domains where the reliability of the estimates in the previous year was less or more

than what was desired.

For D.S.L.P.'s of Class 3 firms, there is no second stage sampling design. Individual shipments are not selected from the files of the D.S.L.P.'s. Instead, aggregated data are collected at the D.S.L.P. level.

## 11. FIELD OPERATIONS

The field operations are different for class 1 and class 2 firms than for class 3 firms. For class 1 and class 2 firms, the operations consist of selecting shipments from the files of their D.S.L.P.'s and of transcribing the characteristics of the selected shipments on coding sheets. For class 3 firms, they consist of obtaining aggregated data over the telephone about their trucking operations.

This section discusses the activities that involve the Statistics Canada Regional Operations staff; namely the training of the Regional Operations project managers, the planning of the collection, the collection at the D.S.L.P.'s of class 1 and class 2 firms, the collection from the class 3 firms and finally the profiling of class 1 and 2 D.S.L.P.'s.

### 11.1 Training of the regional operation project managers

Every year, the Statistics Canada regional operations project managers are trained on all aspects of the survey. The training session is four days long and is conducted during the month of March. It is broken down into two components: an in-class-training and an on-the-job training. The in-class training consists of a series of talks and exercises given by the survey project manager and the methodologist(s). The on-the-job training consists of having groups of three to four people visiting a D.S.L.P. and applying and discussing the knowledge acquired during the in-class training.



## 11.2 Planning of the collection

Having been trained, the regional operations project managers recruit the interviewers and administer a thorough training program. Then the interviewers with the advice of their regional operations project manager schedule their work and plan their itineraries for their visits to D.S.L.P.'s of class 1 and class 2 firms. The itineraries are drawn to avoid unnecessary travel and to achieve maximum productivity. The interviewers mail to the D.S.L.P. officials introductory letters which provide a brief explanation of the survey. Subsequently, the interviewers telephone D.S.L.P. officials for appointments. The collection of the data takes place between May and September for the survey covering the previous calendar year.

## 11.3 Overall description of the collection in the D.S.L.P.'s of Class 1 and Class 2 firms

At the time of the appointment, the interviewer conducts an interview with the D.S.L.P. officials. During the interview, he/she explains the survey, describes the uses of the data, estimates the time required to do the work and asks information about the firm. This information concerns mainly revisions to the names and addresses, changes of ownership, type(s) of document and filing system used and aggregated data about the operations of the D.S.L.P. during the reference year.

The most common types of shipping documents are the probills, bills of lading, load manifests, trip reports, and invoices. A firm may use any combination of these.

The types of filing system include: in complete numeric sequence; in broken numeric sequence; in chronological order; in alphabetical order (e.g. by customer name); by terminals; by commodity type or in no order at all. The documents may even be cross-filed; for example, by serial number and by customer's name. Within a filing system, documents may be kept in a set of file drawers, in sets of binders or shannon files, on shelves, in drawers, or even in books.



The aggregated data about the operations of the D.S.L.P. cover several variables among which are the total transportation revenue earned; the total tonnage carried; the total number of shipments transported; the percentages of each of these three items represented by intercity shipments and the percentages represented by international shipments; the types of commodities carried and the percentage each type represents in the total transportation revenue.

Often the interviewer has a choice of filing systems which provide information on the items needed in the survey. The interviewer assesses the completeness of the various filing systems with regard to the information on the five principal characteristics and on the reference year, and then chooses the system having the smallest under-coverage. However, if two or more systems have the same under-coverage (if any), the interviewer selects the one that includes the smallest number of out-of-scope records or the one that allows out-of-scope records to be removed from the file or not to be counted.

Next, the interviewer selects the sample of shipments as follows. Using the number of shipments reported by the official of the D.S.L.P., he/she gets from a table the corresponding sampling interval and random start. In some instances, the interval and the random start may have already been pre-determined by Statistics Canada Head Office. Next, he/she adds the random start and/or the interval to the document numbers to get the selected shipments in numeric filing system. Otherwise, he/she has to count a number of documents equal to the random start and/or to the interval to get the selected shipments.

Once a shipment is selected, the interviewer transcribes its characteristics. The transcribing operation is often difficult because it can be hard to understand the various documents and the coding used on some documents. This is especially true for the commodity names. The interviewer must avoid the use of brand names, proper names and names which have more than one meaning. The interviewer often has to interpret the information on the documents and to enter on the coding sheets the data in a format that would be accepted by the computer system.

#### 11.4 Overall description of the collection in the Class 3 firms

The interviewer mails an introductory letter two to three weeks prior to any attempt to contact the firm by phone. Subsequently, he/she contacts the official in the firm that is best suited to provide the required information. This may, however, take several phone calls. The interviewer then conducts an interview over the phone.

During the interview, he/she will ask questions similar to the questions asked for class 1 and 2 D.S.L.P.'s. There is a major difference however; no questions are asked about the types of documents utilized and the filing systems used by the firm. Once this first part of the interview is completed, the interviewer proceeds to have the respondent describe his types of shipments. For each type of shipment, the description is to be made in terms of province of origin; province of destination and name of commodity carried. Then the official is asked to report an estimate of the number of shipments, the average weight and the average revenue of each type of shipment.

It is a general subject matter belief that the operations of any given class 3 firms are fairly homogeneous. Therefore each has only a few types of shipments to report. The coverage obtained through this approach is also believed to be acceptable from a user point of view. No testing was done of this hypothesis.

#### 11.5 "Profiling" of Class 1 and 2 D.S.L.P.'s

It sometimes happens that a class 1 or 2 D.S.L.P. cannot provide any documents, does not keep documents suitable for sampling or cannot provide a portion of the shipping documents and that this portion cannot be represented by the available documents. The latter may happen for example when the missing documents represent specific contracts that have been removed for audit purposes. In these cases, the interviewer has to "profile" the missing documents. The profiling consists in having a D.S.L.P. official describe the types of shipments on the missing documents. The profile is similar to the description

of the types of shipments of the class 3 firms with the exception that the precise origin and the precise destination of the shipments (i.e. the village, town, city, etc.) is wanted in this case.

The profiling activity can be long in some D.S.L.P.'s because their operations can be quite extensive. It requires good cooperation from the D.S.L.P.'s official.

## 12. DATA PROCESSING

### 12.1 Manual processing

The completed documents are sent to Statistics Canada Head Office in Ottawa. Upon receipt, the documents are logged in and the identification numbers verified. Two short tasks are also undertaken at this point.

First, a brief scan is conducted to identify and code closings of D.S.L.P.'s, death of firms, out-of-scope firms and abortions. Out-of-scope firms are active firms for which the in-scope revenue is nil for the reference year. Abortions are D.S.L.P.'s for which no information was collected although it was known that the D.S.L.P. had in-scope revenue for the reference year. As examples, a firm found in the field to have earned its revenue 100% from local shipments would be an out-of-scope firm while a single D.S.L.P. firm that refuses to cooperate or is on strike would be an abortion.

Secondly, the profile data of the class 1 and 2 D.S.L.P.'s are examined to determine the number of shipments that should have been transcribed for each reported type of shipments if the documents had been available. These numbers are determined by performing calculations using the total number of shipments covered by the profile, the random start and the sampling interval that should have been used if the documents had been available. These numbers are then coded so that the computer could generate the required number of transcription records for each type of shipments as if transcriptions were obtained.

## 12.2 Data capture

The forms are next sent to data capture. The capture is done on a mini-computer which allows edits and other processing to be performed on-line.

There are many edits performed on the mini. Some edits generate error messages and require corrective action; others generate warning messages that require verification of the entered data and corrective action only if necessary. Some edits consider the validity of each response individually while others consider the relationships between valid characteristics of the same shipment. The operators of the mini-computer are expected to possess subject matter knowledge to perform corrections on-line. Manual imputations are performed when necessary because there is no automated imputation performed for class 3 D.S.L.P.'s.

As part of the other processing, the weight is converted to metric and the rate to \$/100 kilograms. Also, the origin and destination names (i.e. village, town, etc.) if present, are matched against a municipality library to obtain a Standard Geographical Classification (S.G.C.) code, a latitude and a longitude. Whenever there is a nonmatch, the operator is instructed to enter a synonym. Similarly, the commodity name is matched against a commodity library to get a 3-digit Standard Commodity Classification (S.C.C.) code. Whenever there is a nonmatch, the operator uses a synonym or enters "unknown". There is therefore always an S.C.C. code for each shipment. Also, the mini generates the required number of transcription records for each type of shipments from the profile data of the class 1 and 2 D.S.L.P.'s.

Finally, the data are unloaded from the minis and two data sets are created; a data set of shipments of class 1 and 2 firms and a data set of type of ship-data of class 3 firms. The principal difference between the two data sets are that the first one is at the shipment level while the second one is at an aggregated level. Note also that the first one has more variables (e.g. rate, place of origin rather than province of origin, etc.) than the second one.



### 12.3 Main frame edits and imputations

A road distance between the origin and the destination of each in-scope class 1 or 2 shipment has to be obtained in order to be able to provide tonne-kilometres estimates for class 1 and 2 firms. Therefore, the origin S.G.C. - destination S.G.C. pair of each in-scope class 1 or 2 record is matched against a distance library to get a road distance in kilometres between the two locations. Whenever there is a nonmatch, an aerial distance (X) is calculated using the latitudes and longitudes of the origin and of the destination. Then X is converted to a road distance Y using the simple linear regression model

$$Y = a X + b$$

where a and b vary according to 12 regions of origin and 12 regions of destination. The road distance is assigned to the record.

Missing data of partially transcribed shipments of class 1 and 2 firms are also imputed. The imputation technique used depends on the missing variable or the pair of missing variables. Major imputations are performed using fixed relationships between reported figures, unit weight conversion factors and pro-rate tables. An example of a fixed relationship between reported figures is

$$\text{weight} = \frac{\text{revenue} \times 100}{\text{rate}}$$

This relationship can be used to impute weight when revenue and rate are present or revenue when weight and rate are present. Unit weight conversion factors are coefficients determined by unit type (e.g. case, bag, litre, etc.) by S.C.C. code. Knowing the unit and the S.C.C. code of the commodity, the proper conversion factor can be applied to the quantity of units to derive the weight. Finally, pro-rate tables show rates by commodity section, by distance block and by revenue or weight group. These tables are based on the previous years' data modified by incoming valid current-year data. The pro-rate tables are used to calculate the weight when the revenue is present or the revenue when the weight is present.



In cases where too many characteristics of a shipment have to be imputed, the shipment is flagged as not usable.

Expansion edits are subsequently performed. For class 1 and 2 firms, these edits consist of weighting up crudely the number of shipments transcribed, the transcribed revenue and the transcribed tonnage and comparing the results to the total number of shipments, revenue and tonnage provided during the interview by the D.S.L.P. official. Similar edits are performed for class 3 firms. Discrepancies in both cases are followed up.

### 13. ESTIMATION PROCEDURES

For the estimation procedures, it was decided to consider the second stage systematic sampling in the class 1 and 2 firms as simple random sampling without replacement (S.R.S.W.O.R.). This decision was made because first the documents were considered to be in random order and secondly the use of S.R.S.W.O.R. allows the computation of an estimate of the sampling variance.

As the first step of the estimation procedures, weights are calculated. There are first stage and second stage weights for class 1 and 2 records but only first stage weights for class 3 records. In general, first stage weights correspond to the inverse of the probability of selecting of a D.S.L.P. in its stratum and second stage weights correspond to the inverse of the probability of selecting of a shipment in its D.S.L.P. supposing S.R.S.W.O.R. was used. First stage weights are adjusted by computer to reflect the contribution of abortions. No adjustments are made for the closing of D.S.L.P.'s, deaths of firms and out-of-scope firms because they are considered as having generated no shipments. Final weights are attached to each record on the data set of class 1 and 2 firms and on the data set of class 3 firms.

Detailed diagnostic reports are produced. These reports are tables which present the data under various aggregates. They are useful tools to analyse the data and to perform final quality checks.

The data set of class 1 and 2 firms is cleared by discarding out-of-scope shipments. Some types of out-of-scope shipments are shipments to or from the U.S.A.; shipments transported 15 miles or less from origin to destination; shipments which were off-highway; shipments which would be double counted as a result of interlining between road carriers; shipments which would be double counted because they were recorded by household goods movers who are van line agents and by the van lines themselves; shipments which did not generate any intercity transportation revenue; and records which relate to non-transportation services such as storage, packing, equipment rental, labour loading and unloading.

Estimates of revenue, tonnage and tonne-kilometres for the publication are finally generated by summing the weighted data over the appropriate domains. Measures of error such as the coefficients of variation are also provided with the estimates. The coefficients of variation are obtained from the formula derived from the sample design but supposing the systematic sample of shipments is a simple random sample of shipments.

#### 14. USE OF THE DATA AND METHODS OF DATA DISSEMINATION

##### 14.1 Use of data

Requests for estimates yielded from the old survey came from a wide variety of sources. The nature of these requests has also varied a great deal. It is expected that the nature of the demand for data from the new survey will be similar to that in the past.

The estimates have been used extensively to satisfy five main requirements, namely, to measure the volume of domestic trade transported by intercity for-hire carriers provincially and interprovincially; to measure the rate of industrial growth reflected by intercity commodity movements; to provide information on regional development; to assist in transportation studies; and to support the presentation of briefs, submissions and other inquiries to

regulatory authorities and commissions.

One specific use of the data was to define the characteristics of trucking markets using variables such as commodities carried, average lengths of haul and shipment weight. Another specific study examined and analyzed selected aspects of performance of carriers operating in regulated and unregulated environment. The cost behaviour of these carriers was examined by using traffic characteristics such as shipment size and average length of haul.

In the past, special requests for estimates from this survey have come from sources such as government departments concerned with trade, transport regulatory officials at both federal and provincial levels; carriers; university consultants; industry associations; and many other organizations and individuals who share a common interest in transportation.

#### 14.2 Methods of dissemination

The redesigned survey will provide information in three modes similar to the old survey.

First the publication will present the estimates that are generated by the regular system of the survey. Measures of error such as the coefficients of variation will be given with the estimates. Secondly, special requests will be processed subject to cost and reliability constraints. Finally, the data base of shipments generated by the survey may be made available on magnetic tape to selected users subject to constraints of confidentiality.

### 15. FUTURE WORK

As mentioned earlier in this paper, the survey accepts three types of input, one of which is computer tape from selected respondents. This type of input has been found difficult to handle and, although work has commenced on this

subject, progress so far is disappointing. Extensive negotiations are required with the firms to obtain the requested data on tape and then a further analysis is needed to evaluate the data. For reference year 1981, only one tape will be used for a firm which handled about 5 million shipments during reference year 1981. More firms will provide data on tape in subsequent years. Agreements are presently being reached with 5 additional companies for reference year 1982.

Nevertheless, when a firm's computer tape is finally obtained and found to meet our requirements, extensive systems manipulation will still be required to handle the tape. Also, manual interventions will be necessary to handle non matches to the various libraries. Therefore, records will most likely have to be sampled on each tape using the same second stage sampling design as for the sampling of documents of class 1 and 2 firms.

Another area where future work is needed is on having firms themselves sample their documents. As an example, a company could photocopy the pro-bills ending in a given number when the pro-bills are issued, and send the photocopies to Statistics Canada on a monthly basis.

Finally, major efforts will be made to evaluate thoroughly the various phases of the survey and to formulate recommendations for improvements. These recommendations will hopefully be implemented for the 1982 reference year survey.

#### REFERENCES

- [1] Statistics Canada, For-hire Trucking Survey, Catalogue 53-224, Annual.
- [2] "Report on the Findings and Recommendations of the Working Group on the For-hire Trucking Survey Phase I Review"; prepared by the Transportation and Communications Division of Statistics Canada, July 7, 1978.

- [3] "Report of the Interdepartmental Working Group on For-hire Trucking - Phase II Review"; prepared by the Transportation and Communications Division of Statistics Canada, April 1979.
  
- [4] Statistics Canada Motor Carriers - Freight and Household Goods Movers, Catalogue No. 53-227, Annual.
  
- [5] Lussier, R. (1981), "For-hire Trucking Survey: Survey Design", Survey Methodology, Statistics Canada, Vol. 7, No. 1, pp. 74-92.



## THE METHODOLOGY OF THE CANADIAN AIR SCHEDULED INTERNATIONAL PASSENGER ORIGIN AND DESTINATION ESTIMATION SYSTEM<sup>1</sup>

Greg Hunter and Lisa DiPiédro<sup>2</sup>

The Air Scheduled International Passenger Origin and Destination (ASIPOD) estimation system uses the data from two air traffic surveys to produce origin-destination estimates of international passengers. The "assignment technique" is the solution to the problem caused by the non-coverage of non-interlining traffic. The assumptions of the technique are sufficiently questionable to warrant an evaluation of the bias of the estimates. However, major improvements will be made in the new system which will decrease the bias in the estimates. Also, estimates of reliability will be produced. And as a result, knowledge of the strength of the inferences made with respect to air traffic markets from these estimates will be improved in international bilateral air negotiations.

### 1. INTRODUCTION

In 1979, Statistics Canada embarked on a revision of the federal aviation statistics program by inviting Transport Canada and the Canadian Transport Commission (i.e. the two "user departments") to form an interdepartmental revision team. The ASIPOD estimation system is one of several projects in the revision program.

The ASIPOD estimation system uses the data from two air traffic surveys to produce estimates of the number of passengers on scheduled international flights between Canadian and foreign markets for various origin-destination combinations. The first of these two surveys, the revenue passenger origin and destination survey, provides a sample of origin-destination data on international journeys with Canadian carriers on at least one leg of the itinerary.

<sup>1</sup> Presented at the Joint Statistical Meetings of the American Statistical Association in Cincinnati, August 1982.

<sup>2</sup> Greg Hunter, Business Survey Methods Division, Statistics Canada.  
Lisa DiPiédro, Transportation and Communications Division, Statistics Canada.

A coverage problem exists, since no data are available on those international journeys with foreign carriers on all legs of the itinerary. The second survey, the airport activity survey, counts all passengers entering or leaving Canada on all Canadian or foreign scheduled carriers without consideration of the passenger's origin or destination.

This paper first outlines the requirements users have for international passenger origin and destination estimates. Then, the relevant aspects of the two air traffic surveys and the non-coverage problem are presented. And, finally, the paper describes how the ASIPOD estimation system will produce estimates of the number of passengers and the associated coefficients of variation for various international origin-destination pairs for the portions of the international market both covered and not covered by the first survey.

## 2. USER REQUIREMENTS

The users require estimates of international scheduled commercial air service passengers by origin and destination for bilateral air negotiations.

An international scheduled commercial air service is defined to be an operation which is between points in Canada and points in any other country, and which provides public transportation of persons, goods or mail by aircraft in accordance with a schedule and at a toll or charge per unit of traffic. Such a service is referred to as a "unit toll" service.

Before an international scheduled commercial air service can be operated into and out of Canada, some form of formal agreement must exist between the Government of Canada and the government of the second country. The formal agreement between countries may take the form of an interim diplomatic exchange of notes or of a complete negotiated Air Transport Agreement.

International bilateral air negotiations involve officials of the Canadian government from External Affairs, the Canadian Transport Commission, Transport Canada and the Ministry of Industry, Trade and Commerce. Negotiations may last several months or several years.

The routes for scheduled air services are normally the major item for negotiation, but there are many others. Some of the items, or articles written into air transport agreements, may be:

- rights to fly across, or to make stops for non-traffic purposes in, a given territory
- designation of the airline to operate each route
- compliance with laws and regulations of each country, dealing with such issues as entry, clearance, immigration, passports, customs and quarantine
- airworthiness, certificates of competency, and licences
- exchange of statistical information
- tariffs
- transfer of funds
- exemption from taxation of income

In order to negotiate these items, and particularly for an exchange of routes, the negotiating officials must know where air traffic markets are and whether they are growing. Analyses of the costs and benefits to Canada and to foreign countries of various international routes for Canadian and foreign air carriers must be available to the negotiating officials. To ensure that Canada can negotiate a fair market share, the provision of international passenger origin and destination estimates is a necessity. A crude and indirect indication of the value of such data to the Canadian economy is that the revenue<sup>3</sup> generated from all international air routes to and from Canada in 1980 was about 2.3 billion dollars (Canadian).

---

<sup>3</sup> Taken from tabulations internal to the Canadian Transport Commission.

### 3. NATURE OF THE NON-COVERAGE PROBLEM

An underlying assumption of the redesign is that the same basic methodology, as implemented in the existing system, is to be used. As a result, most of the feasibility work involved identifying desirable improvements, ranking their desirability and determining how much could be done within cost and time constraints. This "same basic methodology" provided direction with respect to the calculation of estimates of the number of international passengers, but not to the calculation of estimates of reliability of the passenger estimates.

The target population of the international estimation system is the set of all tickets with an international (i.e. between a foreign country and either Canada or the United States) journey. An exchange program on passenger O & D data is maintained between Canada and the United States, whereby the United States gives Canada those records detailing the complete itineraries of the tickets collected in their survey on which:

- (i) a U.S. and Canadian point is shown in the routing, or
- (ii) a U.S. carrier is recorded as having flown to or from a Canadian point,  
or
- (iii) a Canadian carrier is recorded as having flown to or from a U.S. point.

As a result of this exchange agreement, the expression "foreign" and "foreign (non-U.S.)" are both used in this paper to denote "neither Canadian nor American".

The revenue passenger origin and destination survey collects tickets issued for international journeys, but only major Canadian carriers participate in the survey. Each participating carrier selects a flight coupon on a ticket



with a serial number ending in '0', if that carrier is the first participating carrier to fly on a leg of that ticket. Hence, the survey reports a 10 percent sample of unique flight coupons on which there is at least one participating Canadian or American carrier.

Some information concerning the markets of foreign (non-U.S.) carriers is obtained from the revenue passenger origin and destination survey. For example, if a passenger travels from Ottawa to Montréal with Air Canada, then connects with Air France for Paris, the revenue passenger origin and destination survey will capture the trip because a Canadian carrier participated somewhere in the journey. The Canadian carrier would report the complete carrier and routing detail, including the Air France segment.

The revenue passenger origin and destination survey, however, does not cover coupons with foreign carriers on all legs of an itinerary. An example of such an itinerary would be that of a passenger flying on Air France from Paris to Montréal and then back to Paris on Air France. If this itinerary were the passenger's total journey, this journey would not be reported to the revenue passenger origin and destination survey. However, the itineraries of such passengers are in the target population of international journeys.

This incomplete coverage of the target population is the non-coverage problem for the ASIPOD estimation system. The coverage problem seems to be "non-coverage" as opposed to "undercoverage", since it is not even possible to include a large portion of the universe in the frame.

The existing system takes the revenue passenger origin and destination survey data and the airport activity survey data and applies a method called "the assignment technique" in order to produce total market estimates.

The airport activity survey counts passengers, on a census basis by flight, entering and leaving each Canadian airport. The survey covers all Canadian, American and foreign scheduled carriers, but it does not consider the passengers' initial origin or final destination.



Hence, the airport activity survey provides a count of the total volume of passengers for all carriers in the target population. The assignment technique is a method of estimating the non-coverage volume of passengers and assigning it to origin-destination pairs. However, in order to explain the assignment technique, a somewhat more thorough description of the two air traffic surveys is required.

#### 4. RELEVANT ASPECTS OF THE TWO SURVEYS

The authorizing agency for the two survey programs is the Air Transport Committee of the Canadian Transport Commission, in co-operation with Transport Canada. The data are collected from the air carriers on behalf of the Air Transport Committee by the Aviation Statistics Centre (ASC) of Statistics Canada. Under the authority of the Air Carrier Regulations of the Aeronautics Act, reporting by the carriers on the ASC statements (ie. questionnaires) is compulsory.

The Revenue Passenger Origin and Destination (O & D) statistics are reported to the ASC via Statement 35. The reported data items, among others, include:

- ticket origin and ticket destination
- points of intra- and interlining (i.e. routing)
- carrier on each flight coupon stage

The revenue passenger origin and destination data are submitted monthly by major Canadian unit toll air carriers conducting scheduled passenger services. Since January 1, 1982 the seven Canadian carriers contributing information to this survey have been Air Canada, CP Air, Eastern Provincial Airways, Nordair, Pacific Western Airlines, Air Ontario and Quebecair. The American data are collected by the Civil Aeronautics Board from all certificated United States' air carriers, except helicopter operators and intra-Alaska carriers. The data for the three months of each quarter are combined, and duplicates are

eliminated; so that a file of complete itineraries, Ticket Origin and Destination (TOD) records, is obtained for the quarter.

However, passenger origin and destination statistics are compiled using the Directional Origin and Destination (DOD) concept. The DOD concept can be defined as "points of initial departure and ultimate destination named in the sequence which indicates the direction of travel". DOD's are pieces of itineraries which are broken up such that each component piece defines a reasonably consistent direction. To create DOD's, "open-jaw" and return itineraries, such as "symmetrical" and "circle" itineraries, must be broken into pieces which are essentially one-way trips. To obtain the DOD's, the TOD's are passed through the breakpoint routines. This breakpoint process is automated within the Passenger Origin and Destination System, and involves the calculation of various point-to-point distances within the itinerary and the comparison of these distances to the total itinerary length. As a general rule, itineraries are broken at the farthest point from the origin. Each DOD formed is recycled through breakpoint routines until no further breakpoints can be assigned.

The airport activity data are filed on Statement 32. The relevant items included for each flight are:

- the reporting carrier
- the reporting airport
- the point of origin and final scheduled destination of the flight
- the last station arrived from, for arrivals; or next station departed to, for departures
- the number of deplaned or enplaned revenue passengers

The airport activity data are submitted monthly by Canada's transcontinental

(Air Canada and CP Air) and regional air carriers (Eastern Provincial, Quebecair, Nordair and Pacific Western Airlines), by Norcanair and by all foreign carriers (including American carriers) operating scheduled international flights into and out of Canada. Since January 1, 1982, there have been 10 American and 21 other foreign carriers filing reports for each Canadian airport they served. Each new foreign carrier, granted a licence to serve Canada on a scheduled basis, is automatically included as a participant in the airport activity data collection system.

From the airport activity data, the census traffic flow data are obtained. "Traffic flow" can be defined as a count, over a certain period of time, of the number of persons who are flying on a specific carrier between a Canadian reporting airport and an adjacent point. The adjacent point is called the next stop or the last stop. For the purposes of the assignment technique only the traffic data for foreign (non-U.S.) carriers are input into the system. The data elements extracted from this survey, and used to determine the O & D international markets, are the number of revenue passengers enplaned and deplaned in Canada, the Canadian gateway carrier, and the Canadian gateway. In this survey the concept of "Canadian gateway" is defined to be that reporting airport at which a foreign (non-U.S.) carrier enters or leaves Canada.

However, in the revenue passenger O & D survey, the Canadian gateway for Canadian and U.S. carriers is the first/last Canadian point in the itinerary for a flight entering/leaving Canada. For foreign carriers the Canadian gateway refers to the point inside Canada where the passenger enters or leaves the foreign carrier. Consider the following fictitious example:

Assume that Air France flies Toronto - Montréal - Paris. Some passengers enplaned in Toronto, and some enplaned in Montréal.

Assume also that the matching single crossing DOD's are:

Winnipeg - Air Canada - Toronto - Air France - Montréal - Air France - Paris.

Toronto - Air France - Montréal - Air France - Paris - British Airways - London

The Canadian gateway in the above example would be Toronto because Toronto is the point at which the passengers enter the foreign carrier.

## 5. THE OBJECTIVES OF THE REVISION

The four main objectives, the fourth of which will be discussed in detail in this paper, are as follows:

- (i) to eliminate problems which have been identified in the existing system.

The existing system does not impute for illegible carrier codes on flight coupons, so that "unknown carriers" becomes the third largest carrier in tabulations. Also, there is no check on the coding which indicates whether a carrier is flying to and from airports at which it actually has landing rights. As a result of even existing tabulation requirements, additional edits and imputations, to handle illegible or incorrect data on international flight coupons, are required over and above those required for domestic flight coupons alone.

Other nonsampling errors have been identified, but can not be easily corrected by an estimation system. Examples of such errors are misinterpretation by participating carriers of instructions for selecting flight coupons; systematic errors in serial numbers, used for sample selection, on ticket stock; errors in the carriers' processing systems, etc. The control of these nonsampling errors for which it is not easy to correct is not an objective of this revision.

- (ii) to develop a simple computing system which is easy to use, and produces summary diagnostic information.



Some 625,000 origin and destination records, and some 820,000 airport activity records must be processed annually with minimal manual intervention. Since large volumes of passengers are dispersed across a large number of origin-destination pairs, the diagnostics at each stage of the system must summarize the processing, and still be able to point out potential problems.

- (iii) to tabulate the international passenger estimates, regularly and on an ad hoc basis, in ways which will simplify the analyses undertaken by users.
- (iv) to produce quantifiably reliable estimates of the number of air scheduled international passengers by origin and destination.

Although the reliability of these statistics has been thought to be variable in the past, it has been, in fact, unknown to date. Estimating the reliability of these data will improve the knowledge of the strength of the inferences that can be made from analyses of these data. Inferences made without knowledge of the reliability of the data could actually be quite misleading.

## 6. SOLUTION OF THE NON-COVERAGE PROBLEM

### 6.1 Magnitude of the Problem

As in other surveys, non-coverage is a problem for the ASIPOD methodology, since no sample data are available on the origin-destination patterns of the non-coverage portion of the target population.

The following table gives an indication of the volume of passengers travelling between Canada and nine world areas. (These data are 1979 annual estimates. The world areas are not identified because these data are confidential.)



Table 1 - Estimated Number of International Passengers - 1979

Between Canada and ...	Revenue Passenger Origin and Destination	Non-coverage (percentage of Total in brackets)	Total
World Area #1	116,050	495 (0.4)	116,545
World Area #2	325,200	2,020 (0.6)	327,220
World Area #3	99,010	12,628 (11.3)	111,638
World Area #4	67,200	14,558 (17.8)	81,758
World Area #5	575,010	126,076 (18.2)	703,086
World Area #6	205,180	101,464 (33.1)	306,644
World Area #7	1,221,040	908,876 (42.7)	2,129,916
World Area #8	54,410	56,807 (51.1)	111,217
World Area #9	45,330	57,267 (55.8)	102,597
Total World	2,708,430	1,282,191 (32.1)	3,990,621

From the percentage non-coverage figures, it is evident that the non-coverage problem is a major concern.

The same table as above, but between Eastern Canada and the same nine world areas, would tend to have a higher percentage non-coverage for each world area. Hence, a lower level of geographic aggregation in origin-destination pairs generally implies a higher percentage non-coverage. For example, the non-coverage for Eastern Canada to World Area #7 is 55 percent, compared to the 42.7 percent tabulated for all of Canada to World Area #7 (as above). To clarify the idea that a higher level of geographic aggregation in origin-destination pairs generally implies a lower percentage non-coverage, consider the fact that non-coverage exists only for foreign traffic terminating at Canadian gateways. There is complete coverage of all traffic for which the Canadian end of the origin-destination pair is not a Canadian gateway. Hence, as the level of geographic aggregation becomes higher, more and more interlining traffic is included, and the percentage non-coverage becomes lower.

## 6.2 The Assignment Technique

The assignment technique estimates the non-coverage volume of passengers and then allocates this volume to origin-destination pairs.

From the airport activity survey,  $b$  benchmark counts of passengers entering and leaving Canada at Canadian gateway airports are tabulated by carrier. The value of  $b$ , the number of assignment groups, can be determined as follows:

$$b = 2 \times \sum_{i=1}^g n_i$$

where the '2' accounts for the fact that there is one count each for passengers entering and leaving Canada,

$g$  is the number of Canadian gateway airports, and

$n_i$  is the number of foreign (non-U.S.) carriers with landing rights at the  $i^{\text{th}}$  Canadian gateway airport.

From the revenue passenger origin and destination survey, a corresponding number of inbound and outbound passengers on international DOD's can be tabulated by crossing carrier and Canadian gateway airport. Hence, there are also  $b$  such counts from the sample survey data.

In the first stage of the assignment technique the non-coverage volume,  $A_i$ , of passengers in assignment grouping  $i$  can be estimated as follows:

$$A_i = C_i - (1/f) \times D_i \quad (i = 1, \dots, b).$$

where  $C_i$  is the airport activity census count in assignment grouping  $i$ .

$f$  is the revenue passenger O & D survey sampling fraction (i.e. 1/10).

$D_i$  is the sample number of international passengers in assignment grouping  $i$ .

$A_i$  is, then, an estimate of the number of passengers, carried on the foreign carrier in the  $i^{\text{th}}$  of  $b$  assignment groupings, for which there is no origin-destination information.

The next stage allocates the non-coverage volume to origin-destination pairs. Such pairs in the non-coverage portion are called non-interlining DOD's, since they are DOD's, flown by foreign carriers, which do not interline with a participating Canadian carrier. The assignment technique imputes non-interlining DOD's in the  $i^{\text{th}}$  assignment grouping as follows:

- (i) All of the DOD's contributing passengers to  $D_i$  are identified. (These would be DOD's which match on Canadian gateway, foreign carrier and direction.)
- (ii) The domestic portion of such DOD's (i.e. the portion from the Canadian point to the Canadian gateway city) is eliminated. (The domestic portion of such DOD's would be on a Canadian carrier, and, therefore, would be picked up in the revenue passenger O & D survey. The resultant "truncated DOD's" are, then, non-interlining.)
- (iii) The non-coverage volume, i.e.  $A_i$ , is assigned to the resultant "sample DOD's" in proportion to their original contribution to  $D_i$ .

New DOD records consisting of "assigned passengers" are produced.

The assignment technique assumes that the truncated DOD's are representative of the non-interlining traffic. As a result, some original sample DOD's are receiving more weight than they would in the revenue passenger O & D survey alone. Hence, the estimator,  $\hat{d}_j^T$ , for the total market number of

passengers for the  $j^{\text{th}}$  origin-destination pair can be derived by adjusting the sampling fraction as follows:

$$\hat{d}_j^T = (1/f_j) \times d_j$$

$$\text{where } f_j = \frac{d_j}{(d_j/f) + a_j} \quad (1)$$

and where  $f_j$  is the adjusted sampling fraction associated with the  $j^{\text{th}}$  origin-destination pair, from the revenue passenger O & D sample survey.

$d_j$  is the sample number of international passengers in the  $j^{\text{th}}$  origin-destination pair, from the revenue passenger O & D sample survey.

$a_j$  is the number of passengers assigned to the  $j^{\text{th}}$  origin-destination pair.

Note that  $a_j$ , according to point (iii) above,

$$\text{is } a_j = \frac{d_j}{D_i} \times A_i \quad j \in i$$

$$\text{where } \sum_{j \in i} d_j = D_i$$

$$\text{and } \sum_{j \in i} a_j = A_i$$

### 6.3 An Example of the Assignment Technique

A simple example will illustrate how the assignment technique works. Assume that the airport activity data from British Airways indicated that 120 passengers enplaned at Montréal and went to London. Therefore,  $C_i = 120$ . Assume that the only two DOD's for this assignment grouping from the revenue passenger origin and destination survey are:

<u>Sample Number of Passengers</u>	<u>Estimate of the Number of Passengers</u>	<u>DOD's</u>
1	10	YWG-AC-YMX-BA-LON-LO-WAW
2	20	YYZ-CP-YMX-BA-LON-LH-HAM

where the codes are to be interpreted as:

<u>Code</u>	<u>Denotes</u>
YWG	Winnipeg
YMX	Montréal (Mirabel)
LON	London
WAW	Warsaw
YYZ	Toronto
HAM	Hamburg
AC	Air Canada
BA	British Airways
LO	LOT
CP	CP Air
LH	Lufthansa

Therefore,  $D_i = 3$ , and  $A_i = 120 - (1/.1) \times 3 = 90$ .

The truncated DOD's, the proportion of their contribution to  $D_i$ , the resultant number of assigned passengers and total market estimates are then:

<u>Truncated DOD's</u>	<u>Proportions</u>	<u>Assigned Passengers(<sup>a</sup>j)</u>
YMX - BA - LON - LO - WAW	1/3	30
YMX - BA - LON - LH - HAM	2/3	60



#### 6.4 Shortcomings of the Assignment Technique

The assignment technique has some recognized shortcomings.

The basic assumption of the assignment technique is that the truncated DOD's are representative of the non-interlining traffic. Consider the hypothetical example above in order to determine whether this assumption is intuitively reasonable. The assignment technique presumes that passengers flying on a particular foreign air carrier and originating in Montréal would have the same ultimate destination as passengers originating in Winnipeg or Toronto who fly through Montréal. Hence, the travel patterns of ethnic communities, for example the Polish and German communities in Toronto and Winnipeg respectively, might be used to impute for travel patterns of the more predominantly French communities in Montréal. And, in fact, the assumption that interlining and non-interlining travel patterns are the same was proven empirically to be suspect in a pilot test (Rosen and Conroy (1977)) conducted by the Canadian Transport Commission in 1977. Therefore, there is not only some intuitive but also some empirical evidence against the basic assumption of the assignment technique.

The accuracy of the estimates of the number of passengers by origin-destination pair is jeopardized by any violations of the assumption that truncated DOD's are representative of non-interlining traffic. Large volumes of passengers are allocated to origin-destination pairs, as was seen in Table 1 above, based on a small "effective" sample size. For example, the "effective" sampling fraction of the Canada - World Area #7 market is, not 10% as in the domestic survey, but 5.7% (ie.  $(100\% - 42.7\%) \times 10\%$ ), because of the non-coverage of non-interlining traffic. Hence, a smaller than 10% sample of DOD's is used to allocate a large volume of passengers. Violations of the aforementioned assumption, then, would cause a potentially large bias in the estimates.

Since airport activity census counts are used as benchmark figures for traffic volumes, their accuracy is very important. Although no evaluations have been

undertaken to investigate the magnitude of bias from nonsampling errors in the airport activity census counts, aviation statistics economists feel that this is a survey in which such errors would be small. There are currently, however, ongoing discussions with the major Canadian air carriers on how the reporting requirements of government agencies can be minimized. As a result of these discussions, the airport activity survey could become a sample survey. If a sample is to be designed, the accuracy of the activity counts of gateway airport passengers would be an important design consideration.

The assignment technique also assumes that all of the non-coverage is accounted for by the non-interlining traffic. An interesting way of validating this assumption would be to compare, for the same reference period, the airport activity census count for a Canadian air carrier to the analogous estimate from the origin-destination sample survey. This analogous estimate would be the sum of all passengers on DOD's with the same Canadian gateway airport and crossing carrier. It would be necessary to be able to determine whether differences in the estimates were ascribable to differences in the concepts of the two surveys, and if so, and whether these differences have been taken into account in the ASIPOD estimation system. If such differences have not been accounted for, then it could be that there is a problem with the assumption that all of the non-coverage is accounted for by non-interlining traffic.

Many of these shortcomings should be investigated. However, there are major problems in the existing system, and no alternative solutions, which are superior to the same basic methodology of the assignment technique and which can be implemented within time and cost constraints, have been found. Furthermore, the improvements in the estimates of the number of passengers by air traffic market in the new ASIPOD system will be substantial.

The use of the assignment technique to estimate for international markets is an innovative solution to a large problem. It does not completely solve the

non-coverage problem, but it is a major step in the right direction, as will be the production of estimates of the variance of the estimates. These variance estimates should take account of the assignment technique and its assumptions, and, at the same time, give a meaningful measure of the reliability of the DOD estimates.

## 7. ESTIMATION OF VARIANCE

The estimator of the variance of the international origin-destination estimates is a simple extension of the variance estimator for the revenue passenger origin and destination survey.

### 7.1 Variance of the Estimate of Interlining Traffic

The development of the estimator of the variance of the revenue passenger origin-destination estimates is dependent upon the way in which tickets contribute passengers to origin-destination pairs (i.e. the domains of interest). Recall that each ticket is selected with probability 0.1, and that each ticket may be broken up into several segments or DOD's. Each ticket may contribute 0,1,2, etc. passengers to a given domain of interest. For example, the itinerary

YWG - AC - LON - BA - YYZ

would be broken, via the breakpoint routines, into the two DOD's

YWG - AC - LON  
and  
LON - BA - YYZ.

Consider the inbound plus outbound estimates which are total passenger figures, independent of direction. For such estimates this ticket would add passengers to, among others, the following domains of interest:

<u>Domains of Interest</u>	<u>Passenger Count</u>
Winnipeg - London	1
Toronto - London	1
Canada - London	2
Canada - Europe	2
Eastern Canada - Europe	1
Western Canada - Europe	1

Note that the number of passengers per ticket depends on the geographic level of aggregation, and, therefore, on the particular origin-destination estimate (i.e. domain of interest).

The estimate,  $\hat{d}$ , of the number of passengers from the revenue passenger origin and destination survey can be developed for a particular domain of interest as follows:

$$\hat{d} = \frac{n}{\sum_{i=1}^n} \frac{x_i}{f}$$

where  $x_i$  is the number of DOD's belonging to the domain of interest on the  $i$ th ticket.

$n$  is the number of sample tickets selected, and

$N$  is the population number of tickets in the revenue passenger origin and destination survey.

$f$  is the sampling fraction, i.e.  $n/N = .1$

The revenue passenger origin and destination survey sample is effectively a 10% simple random sample because

- (i) the selection of coupons with serial numbers ending in the digit '0' produces a systematic sample, and

- (ii) there is no cycle associated with the distribution of tickets which would cause a relationship between the survey estimates and the last digit of the serial number.

The estimate of the variance can be written, then, as

$$\text{var}(\hat{d}) = N^2 \frac{1}{n} (1 - f) v_s$$

$$\text{where } v_s = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{and } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i ,$$

and the coefficient of variation, as

$$cv(\hat{d}) = \sqrt{\hat{\text{var}}(\hat{d})/\hat{d}} .$$

Note that  $\hat{\text{var}}(\hat{d})$  can also be written as

$$\hat{\text{var}}(\hat{d}) = \frac{(1-f)}{f^2} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

where  $n$  is assumed to be sufficiently large for  $n/(n-1)$  to be approximately equal to 1.

## 7.2 Variance of Total Market Estimates

The method for calculating total market coefficients of variation has to recognize that the assigned data are a function of the sample data. In other words different samples will produce different assigned data and, thereby, different values of the total market estimate. The re-use of certain portions



of the sample has an effect on the sampling distribution of the estimates.

The method which will be used adjusts the sampling fraction from 10% to be the percentage for which sampled records for a particular domain of interest are actually accounting. Since the use of the assignment technique is a given, it has to be assumed that truncated DOD's from the revenue passenger origin and destination survey are representative of the non-interlining traffic. The measure of reliability will be a measure of the precision of the DOD estimates, only to the extent to which this assumption is valid. As a result, it is reasonable to adjust the weights of the sample DOD's to take into account the non-interlining DOD's. The sampling fraction, then, for the estimation for the  $j^{\text{th}}$  domain of interest would be  $f_j$  as developed in equation (1) above.  $f_j$  would replace  $f$  in the formula for  $\hat{\text{var}}(\hat{d})$  in equation (2) above in order to yield the formula for the variance of the total market estimates:

$$\hat{\text{var}}(\hat{d}_j^T) = \frac{(1-f_j)}{f_j^2} \sum_{i=1}^n (x_i - \bar{x})^2$$

And, the coefficient of variation for the total market estimate would be

$$\hat{\text{cv}}(\hat{d}_j^T) = \hat{\text{var}}(\hat{d}_j^T) / \hat{d}_j^T$$

Note that  $f_j < .10$  for  $a_j > 0$ . This means that less than a 10% sample is achieved when it is necessary to include assigned data in a total market estimate. Hence, by using  $f_j$  instead of .10 in the expression for the variance of a DOD estimate, the coefficient of variation is adjusted to relate to the total market estimate.

This method gives credit to the use of sample data in the assignment technique; but it is dependent, as is the determination of the estimates themselves, upon the assumption that truncated DOD's are representative of non-interlining traffic.

## 8. FUTURE CONSIDERATIONS

Earlier, the need for data on air traffic markets in international bilateral air negotiations was explained. The exchange of statistical information is one of the negotiable articles in air transport agreements. Currently, agreements for the exchange of statistical information exist with several countries. The concepts and quality of the data from some of these countries indicate that these data could be used in the ASIPOD estimation system. Such data would provide sample information on the non-coverage portion of the international universe of tickets. Feasibility work is currently underway to determine whether the number of countries for which these data can be used would improve the accuracy of a sufficient number of estimates to justify the expansion of the ASIPOD system to use "exchange data".

## 9. ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the Transportation and Communications Division and the Business Survey Methods Division of Statistics Canada, and in particular Robin Dunn, for their many helpful comments during the development of this paper.

## REFERENCES

- [1] Burchell, J.M.: "International Air Passenger Origin/Destination Project"; presentation to the Transportation and Communications Division of Statistics Canada, March 1981.
- [2] "International Air Passenger O & D Statistics System - User Requirements Specifications"; prepared in the Aviation Statistics Centre (ASC) of the Transportation and Communications Division of Statistics Canada, revised March 1981.
- [3] "International Air Passenger O & D Statistics System - Feasibility Study Report"; prepared in the ASC, January 1981.

- [4] "International Air Passenger O & D Statistics System - Requirements for Functional Specifications (Concepts)"; prepared in the ASC, revised September 1981.
- [5] Rosen, F.G. and Conroy, P.F.: "A Test of the Assignment Technique Based on a Survey of International Air Travellers at Montreal and Toronto"; a confidential report prepared for the Passenger and Aviation Economics Directorate of the Research Branch of the Canadian Transport Commission, February 1977.
- [6] Carpenter, R.: "Specifications for Estimates of the Reliability for the Air Scheduled International Passenger Origin and Destination Estimates"; prepared in the Business Survey Methods Division of Statistics Canada, revised January 1982.

## SOME ASPECTS OF QUALITY OF CANCER MORTALITY AND INCIDENCE STATISTICS

D. Binder and A. Malhotra<sup>1</sup>

Statistics Canada, Canada's central statistical agency, has been compiling national mortality statistics, including those on cancer mortality since 1921. Also, cancer incidence data are available from 1969.

The data quality of these files may be assessed in a variety of ways. Ratios of cancer mortality to incidence give some information on coverage errors. Micro-data matches between incidence and mortality files give an indication of misclassifications. As well, multiple registrations for cancer incidence may be duplicates. Completeness and availability of data items are also important for special studies.

In this paper, the feasibility of using these measures of data quality and the implications of these measures are discussed.

### 1. INTRODUCTION

Population based cancer statistics are the basis of epidemiological research into the distribution and determinants of cancer and underlie health programmes for the prevention, diagnosis and treatment of cancer. Statistics Canada, Canada's central statistical agency, compiles two such types of data on cancer.

1. National mortality data which are based on reports from provincial vital statistics registration systems. These data date back to 1921.
2. National cancer incidence data which are based on notifications from provincial cancer registries. This data series was established in 1969.

---

<sup>1</sup> D. Binder, Institutional and Agriculture Survey Methods Division, Statistics Canada and A. Malhotra, Health Division, Statistics Canada.

Good statistics which provide reliable information on risk differentials depend on completeness and accuracy of cancer registration and comparability of the data between different registration areas and time periods. Cancer incidence and mortality data as reflections of the true cancer risk each have their own merits and limitations.

Cancer incidence data are particularly suitable for the epidemiological study of cancer because they provide information on all cancers, not only those that are fatal, because they can provide an early warning of emerging problems and because the diagnostic information is usually detailed and of a high quality. For example, the publication *Cancer Incidence in Five Continents* [1] emphasizes the role that international comparisons of cancer incidence play in yielding clues about the causes of cancer in spite of certain well known limitations of the data. These limitations include difficulty in achieving complete registration of new cancer cases and differences between registries in the extent to which this is achieved. Major factors influencing coverage are the number and types of data sources used, how active case finding is, the length of time a registry has been in operation and whether or not the reporting of cancer is a legal requirement in the registration area. Canadian cancer registries are quite heterogeneous in their data collection methodology but all attempt to follow international [2] as well as national [3] guidelines for the standardized recording of cancer incidence data. Differences in sources and techniques of registration not only influence coverage but also other aspects of data quality such as detail of socio-demographic and geographic information that is provided. Also, cancer incidence data are sensitive to such factors as mass screening programmes which result in the inclusion of previously undiagnosed prevalent cases.

The editors of *Cancer Incidence in Five Continents* use and discuss a number of indices which may be useful in assessing completeness of registration and reliability of the data [4]. These include cancer mortality-incidence ratios as indicators of completeness of registration (see Sections 2.1 and 3.1 of this paper).

The reporting of deaths is a legal requirement in most developed countries so that coverage error is assumed to be small. Known and suspected limitations



of mortality data for purposes of epidemiologic cancer research include a lack of information on non-fatal cancers, less precise diagnostic information and more frequent misclassification due to assignment and coding of underlying cause of death resulting in less accuracy compared with the diagnostic information in cancer incidence statistics.

A quality assessment of vital statistics [5] which was undertaken as a pilot study gives some indication of the quality of Canadian mortality data (all causes), particularly on error rates in the coding of underlying cause of death. This error rate was 7.2% in the data year 1976. About two-thirds of the errors involved the first or second digit of the 4-digit cause of death code. Variation in the error rate by specific causes of death was not investigated.

This paper is concerned with assessing the feasibility of measuring certain aspects of quality of the two cancer data files at Statistics Canada which are used in epidemiological studies, namely the cancer incidence file and subset of records from the mortality file with an underlying cause of death of cancer.

The aspects of data quality selected for investigation were:

1. Completeness of registration of new cancer cases through a comparison of cancer mortality with cancer incidence in the same period. This comparison is a crude but commonly used indicator of completeness of registration.
2. Consistency of assignment of diagnosis and cause of death codes through matching individual records on the two files.
3. Availability and completeness of data items through an analysis of how often valid values are present on the files.
4. Registration of multiple primary cancers on the incidence file through a matching of records within the file.

The period covered by the study is 1969-1978, the period for which cancer incidence data are available. Ontario was excluded from all investigations because the National Cancer Incidence Reporting system includes data for this province for 1969-1971 only<sup>2</sup>.

A discussion of the approach taken in the investigations is contained in Section 2 of this report. In Section 3 the findings are discussed.

## 2. DESCRIPTION OF MEASURES

In this section we describe some methods for studying the data quality of the cancer mortality and incidence files.

### 2.1 Mortality-Incidence Ratios

In order to study relative rates of undercoverage among cancer incidence registrations, we consider the ratios of deaths to incidents of cancer. Since the Mortality System registers all deaths in Canada by cause of death and the concept of the National Cancer Incidence Reporting System (NCIR) is to register all new incidents of primary malignant neoplasms<sup>3</sup>, if the two registration systems were of the same quality for all reporting registries, one would expect that the ratio of mortality rate to the incidence rate for a particular site would be fairly consistent within a population of given age and sex over sufficiently long periods of time. (We compute these ratios for deaths and incidents over 5-year and 10-year periods to reduce the effect of the time lag between the reporting of a cancer incident and death). Inconsistency of this ratio would arise if any of the following rates differ across reporting registries:

- (a) rates of survival or sudden increase rate of incidence
- (b) mortality rates from other competing risks,
- (c) rates of error in coding of underlying cause of death,

---

<sup>2</sup> Ontario developed a passive registration system which makes use of reports on cancer patients made for other purposes. Data for recent years are currently being prepared by the province.

<sup>3</sup> By exception, metastatic cancers are registered when the primary site is unknown.

- (d) rates of error in classifying cancer site for new cancers,
- (e) rates of under-reporting or over-reporting of cancer incidents.

If the mortality-incidence ratios are grossly different for most sites, then differing rates in (a) and (b) can be discounted. With respect to the error rates in (c), studies on the coding error for underlying cause of death have yielded error rates of less than 10% [5]. In an unpublished report it was found that these error rates could vary from 3% to 18% across reporting registries. However, the observed differences in the mortality-incidence ratios (see Tables 1 and 2) cannot be completely explained by these error rates. Also, since about 90% of the registrations of new cancers are confirmed histologically, the error rate in (d) would be small. Therefore, these mortality-incidence ratios do give an indication of the coverage error in the NCIR System.

Concentrating on those sites with leading diagnosis count (excluding skin cancer), based on the NCIR file, for each sex, we show in Tables 1 and 2 the national mortality-incidence ratios as well as provinces with largest and smallest ratios, for two five-year periods, broken down by age-groups. We omit Prince Edward Island from consideration because the number of observed events is too small for valid comparison.

## 2.2 Matching Mortality and Incidence Records

In order to assess the feasibility of evaluating errors in cause of death classification, or errors in cancer site classification, a sample of records can be selected from either the Mortality File or Cancer Incidence File and then the other file can be searched for matching records. The manual search does not guarantee that all true matches will be found, and, in fact, the rate of successfully matching may be different across reporting registries, because the level of detail of matching variables can vary from one registry to another. (See Section 2.3 for a study on availability of data).

Records with malignant neoplasms of the lung or bronchus (ICDA-8<sup>4</sup> is 162.1) for the years 1969-1978 were selected as starting points from both files. This choice was based on considerations of high incidence, high mortality and short survival times so that conditions were favourable for finding matching records on both files. In spite of this, because of the time difference between diagnosis and death, it is true that cancer deaths in the earlier years and newly diagnosed cancers in the later years are less likely to have a corresponding record on the matched file. The analysis of the results would be improved in future studies if the sample design controlled for year of death and year of diagnosis.

Two independent samples were selected: one from Mortality File and the other from the NCIR file.

#### 2.2.1 Mortality to Incidence

All deaths from cancer between 1969 and 1978 should have, at least conceptually, a corresponding record on the NCIR system. Noteworthy exceptions to this rule are that the cancer was first diagnosed in Ontario or outside Canada or that the cancer was first diagnosed before 1969. Besides the exceptions, a lack of a corresponding record on the NCIR system is an indication of under-coverage. This, of course, assumes that the underlying cause of death in the Mortality File is error-free.

Therefore, if a sample of deaths from cancer are selected and matched to the NCIR system, we have a number of possible outcomes:

- (a) no matching record is found,
- (b) a match was found with a record having a different cancer site,
- (c) a match was found with a record having the same cancer site.

If no match is found, this is an indication of under-coverage, or that the cancer was first diagnosed in Ontario or outside Canada or prior to 1969, or that the death was not really a death due to cancer. Alternatively, as

---

<sup>4</sup> International Classification of Diseases, Adapted for Use in the United States, Eighth Revision.



previously mentioned, the matching process itself is not perfect. If a match is found, but the records have different cancer sites, this may be an indication of an error on either the Mortality File or the NCIR System. As mentioned previously, it is generally believed that the NCIR system yields the more accurate disease classification, because of the high rate of histological confirmation.

A small scale study was undertaken to measure this phenomenon. A random sample of 56 records with underlying cause of death reported as a malignant neoplasm of the lung or bronchus (ICDA-8 is 162.1) was selected from each of the provinces except Ontario yielding a total sample of 504 records. Only deaths between 1969 and 1978 were selected.

The national rate of successful matches was 82.3%. The rates varied between 73.2% and 96.4% across the nine provinces. Among those with successful matches, 92.5% had the same 4-digit ICDA-8 classification. These rates varied from 74.4% to 100.0%. For those provinces with the lowest and highest rates of matches with the same ICDA-8 classification, we give in Table 3 the breakdown of the observed disease classifications.

### 2.2.2 Incidence to Mortality

A sample of records from the NCIR System was also selected and matched to the Mortality File. Fifty-six records with malignant neoplasms of the lung and bronchus (ICDA-8 is 162.1) from each of the nine reporting registries were randomly selected yielding a total sample of 504 records. The matched records were then checked for underlying cause of death on the complete Mortality File for 1969 to 1978. We did not check the cause of death on the original death certificate for this study, although this would be feasible for future studies.

The outcomes from this manual match may be classified as follows:

- (a) no matching record is found,
- (b) a match was found with a cause of death other than cancer,
- (c) a match was found with a cause of death being cancer but not cancer of the lung or bronchus,



(d) a match was found with the same cause of death.

If no match is found, this is an indication of the inadequacy of the matching process, unless death occurred after 1978 or outside Canada or the person is still alive.

A match found with a different underlying cause of death is an indication of one of the following:

- (a) a competing risk took precedence,
- (b) cancer was a contributing cause of death but not the underlying cause of death or
- (c) the underlying cause of death on the Mortality Data Base was incorrect, or the cancer site was incorrect on the NCIR system (the latter being assumed less likely).

The average rate of successful matches was 69.4%. The rates varied between 55.4% and 80.4% across the nine provinces. Among those with successful matches, 92.0% had the same 4-digit ICDA-8 classification. These rates varied from 85.3% to 100.0%. For those provinces with the lowest and highest rates of matches with the same ICDA-8 classification, we give in Table 4 the breakdown of the observed cause of death classifications.

### 2.3 Availability and Completeness of the Data

One simple measure of the quality of the data files is the relative frequency of valid data for specific items. For the National Cancer Incidence Reporting System and the cancer deaths on the Mortality File, we concentrate on the following items:

- date of birth (day, month, year)
- age
- place of birth
- county and subdivision of residence

We chose these items to exemplify how easy or difficult it would be to match records from other files (e.g. Section 2.2), or to create special tabulations, such as small area statistics. For each item we classify the data as being

valid or invalid. Besides blank values, invalid data would arise when alphabetic characters are found in a numeric variable or the numeric value is out of range. We have aggregated the relative frequencies into two five-year groupings (1969-1973 and 1974-1978) so that we can see whether the quality has changed significantly in the later years.

In Tables 5 and 6 we report the national averages for the two data bases as well as show the values for the provinces with largest deviation from the national average. For the Mortality File, we give the results only for cancer deaths in the nine provinces outside of Ontario so that the comparison with the NCIR system is more meaningful.

#### 2.4 Multiple Registrations on the Cancer Incidence System

The concept of the National Cancer Incidence Reporting System is to register all new incidents of malignant neoplasms. An individual should be registered more than once when multiple malignant neoplasms develop. To avoid duplicate registration of the same incident or duplicate reporting of patients registered in more than one province, all provincial cancer registries follow routine procedures. In spite of this, duplicate reporting of the same cancer incident may occur. In order to evaluate the extent of the duplication, we searched for records which are likely duplicates. The search was nowhere near exhaustive, so that the number of potential duplicates found is an underestimate. Of the 457,158 records, we removed the records with invalid surnames or years of diagnosis. For those with missing birthyear, we calculated the birthyear from the age when available. We also removed skin cancer records (ICDA-8 is 173) since this is known to have multiple occurrences.

Of the remaining records, we found those cases where all the following occurred:

- birthyear or calculated birthyear matched exactly,
- surname matched exactly,
- first four letters of first given name agreed,
- the three digit ICDA-8 code agreed.

Of these records, we identified multiple registration as follows:

- year, month and day of birth was present and agreed, or
- day of birth was not present on at least one record but month of birth agreed.

We also manually verified all groups with at least 3 individuals where the month and day information did not agree and all groups of 2 individuals where the month or day information was missing on at least one record.

In all, this resulted in identifying 6113 records which were potential duplicates. These records correspond to 5947 individuals. (Note that some individuals were duplicated more than once.) We did not make the judgment as to whether these were legitimate multiple registrations or actual duplicates.

For each 3-digit ICDA-8 value, we show in Table 7 the breakdown of these potential duplicates according to whether the records came from the same reporting registry or different registries, as well as how many potential duplicates have the same fourth digit of the ICDA-8 classification.

### 3. DISCUSSION

#### 3.1 Mortality - Incidence Ratios (Tables 1 and 2)

Ratios of cancer mortality to incidence can provide an indication of completeness of registration. The ratios will vary with cancer site (the highest ratios occur for sites with the lowest survival), age and sex for all registries. However, if a comparison of the ratios for different registries shows major differences within a given site, sex and age group, differences in completeness of registration of new cases of cancer must be suspected. A higher ratio, which means a higher proportion of deaths compared with newcases in the same period may indicate less complete registration of new cases.

In both time periods there were two registries which consistently had the highest ratios for all sites combined and for most of the major sites shown. There is little doubt that these high ratios do reflect underregistration of new cases - the registries are the only ones which do not use death

notifications as one of their sources of registration. In addition, one of these registries uses only a single data source, hospital reports, to register cancer cases. This registry had previously reported the results of a special study which showed that it was receiving notifications for only an estimated 70% of new cancer patients admitted to hospitals up to the end of 1976. Following this, major changes were made to improve the notification system. Since 1977 the registry has been reporting a higher number of cancer cases which is reflected in a marked reduction in mortality-incidence ratios.

All other Canadian cancer registries use multiple sources of registration which is considered essential to achieve good coverage and which could also have a positive impact on the completeness and quality of individual data items. The completeness of reporting of data items was examined in this study (see Subsection 2.3). However, it turns out that the one registry that uses only one data source actually ranks quite highly in terms of completeness of information for many data items.

A possible drawback associated with using multiple sources of registration is that duplicate registration may result. However an analysis of multiple registrations for the same individual and the same cancer site does not bear this out. In general, registries using a larger number of different sources of registration do not have more multiple registrations for the same site than registries using fewer sources of registration.

Cancer mortality-incidence ratios for the other six registries were more similar to each other. For these registries there was no consistent pattern of one registry always having higher or lower ratios for all sites and both time periods.

There are many factors that can influence variations in the observed ratios by cancer site. Factors which tend to result in less complete registration of new cases and therefore higher mortality-incidence ratios includes difficulty in diagnosing the cancer (e.g. in deep-seated organs) and lack of access to specific data sources (e.g. haematology reports confirming a diagnosis of leukaemia).



Factors which may lead to overregistration of new cases and lower ratios are mass screening programmes (which may lead to the inclusion of prevalent cases, especially, for slow-growing tumours), duplicate registration, inclusion of in-situ cases and inclusion of latent cancers discovered only at autopsy (this particularly affects cancer of the prostate). In addition, differences in the accuracy of assignment of diagnosis or cause of death may lead to artefactual differences. For example, death certificates may state "cancer of the uterus, unspecified" or "leukaemia, unspecified" as the cause of death whereas a cancer registry will often have more precise information and will assign more precise codes [6]. An analysis of mortality-incidence ratios at the level of the more detailed diagnosis would therefore show gross discrepancies.

Of the leading cancer sites that were examined for males, cancer of the lung and stomach were associated with the highest mortality-incidence ratios for all registries but interprovincial variation in the ratios was greatest for cancer of the colon (excluding rectum) prostate and bladder. Use of cancer incidence data in studies designed to identify differences in cancer risk by geographic area (province) would therefore be more reliable for the former two cancer sites.

For females, of the leading sites examined, cancer of the colon and ovary had the highest ratios for all registries. Interprovincial variation in the ratios was greatest for cancer of the uterus and cervix uteri as well as for cancer of the colon. For purposes of interprovincial comparisons, incidence data for breast cancer and cancer of the ovary would therefore be more reliable.

In the case of the sites of cancer of the uterus (other than cervix) and cancer of the cervix, there are large interprovincial variations in the ratios if the sites are considered separately. This variation is greatly reduced if the two sites are combined, suggesting that there are differences in the accuracy of diagnosis and cause of death assignment for these sites.

The site-specific mortality-incidence ratios were examined for major age groups. The highest ratios consistently occur at older ages (65 and over) for



all registries and all sites shown. This is as expected, since the risk of death increases with age so that proportionately more death than new cases occur at older ages. It is also recognized that diagnosis and registration of cancers in older persons is generally more difficult. However, the relative increase in the ratios at older ages is much greater for the registries which have the highest average ratios to start with. This indicates that while all registries may have some difficulty in registering older patients, under-coverage of the older population is greater for registries which in general have less complete registration systems.

The Canadian data therefore lend support to recommendations made by the International Union against Cancer (1970) and the International Agency for Research on Cancer (1976) and the reiteration of this recommendation in a recent paper by Doll and Peto [7] that "reasonably reliable comparisons of cancer incidence are obtained only if comparisons are limited to men and women in middle life".

### 3.2 Matching Mortality and Incidence Records(Tables 3 and 4)

Since Statistics Canada is responsible for managing both the cancer incidence and the mortality data files, it is possible to compare reports for individuals who are listed in the two separate data files to verify the reported information.

In this part of the investigation of data quality, the accuracy of assignment of diagnosis and cause of death was of particular interest. Within the scope of the study it is only possible to describe the results - the reasons underlying the discrepancies found remain unknown. However, it is felt that the findings are revealing and do indicate, in the case of the particular cancer selected for analysis, lung cancer, that agreement on diagnosis between the two files is generally high, over 90%.

The study also indicates that a larger scale match would be feasible to assess the accuracy of diagnosis and cause of death codes. Of course, if a larger scale study were based on a sample, it would be preferable to stratify the sample by year of diagnosis or year of death. In theory, if computerized

matching techniques were used, this type of analysis is possible for all cancer sites. If such an undertaking were to be supplemented with, for example, studies on accuracy of coding of cause of death and diagnosis in the field, such as described in two U.S. reports [8] [9], interpretation of epidemiological research findings would be facilitated.

### 3.2.1 Mortality to Incidence File Search

Of the sample of 504 death records with an underlying cause of death of lung cancer from 1969-1978, 415 (82%) corresponding records on the cancer incidence file for the years of diagnosis 1969-1978 were found<sup>5</sup>. The rate of unsuccessful matches varied from 3.6% to 26.8% across the nine provinces. This rate is influenced by four main factors: (a) that the cancer was first diagnosed prior to 1969, (b) that it reflects underregistration of new cases, (c) that identifying information was not adequate to permit matching of records, or (d) that the cause of death code is incorrectly given as cancer. There were insufficient data to allow assessment of the relative contribution of each of these factors.

Of the 415 death records for which a corresponding record was found on the incidence file, there was agreement on the diagnosis, primary cancer of the lung, in 92.5% of cases. There was 95.2% agreement that a cancer of the respiratory system was present. The small number of remaining records, had diagnoses for sites other than respiratory cancer on the incidence file. It is generally accepted that the diagnostic information on cancer registry files is more accurate than the cause of death information on death certificates. However, given the scope of this study it is not possible to determine if misclassification on either of the files (or perhaps the fact that a lung cancer was first diagnosed prior to 1969 followed by a subsequent registration for another primary cancer) account for the disagreement. Across the provinces, the rate of agreement on diagnosis varied from 74.4% to 100%.

<sup>5</sup>

In six cases more than one corresponding record for the same individual existed on the incidence file. Only one of these records was counted as a successful match.

Interestingly, for the province with total agreement on diagnosis there was also the highest success rate (96.4%) of locating corresponding records on the two files. This could possibly indicate close liaison between the provincial vital statistics office and the cancer registry. In the reverse match of a sample of cancer incidence records to mortality records (described in Section 3.2.2) it was the same province that had the highest rate of successful matches as well as complete agreement on diagnosis.

### 3.2.2 Incidence to Mortality File Search

The reverse search using the incidence file for the years 1969-1978 as a starting point and attempting to locate a corresponding record on the complete mortality file for the same period was successful for 69.4% of the selected sample of 504 records with a diagnosis of primary lung cancer. The rate of unsuccessful matches was higher than in the match from mortality to incidence records for all provinces and varied from 19.6% to 44.6%. Possible reasons for not finding a match include (a) that the patient was still alive at year end 1978, or (b) that identifying information was not adequate to permit a match. It is in general less likely that one will find a corresponding record in the search from incidence to mortality file since some persons diagnosed to have lung cancer do survive this whereas all persons who die from lung cancer should be registered as new cases either prior to death or at time of death.

For the records that were successfully matched there was agreement that the diagnosis was a primary lung cancer in 91.4% of cases, a rate very similar to that found in the reverse comparison. The samples for the two comparisons were chosen independently so the consistency of the findings concerning agreement on diagnosis is reassuring. Of the remaining cases, 4.3% had the underlying cause of death classified to cancer sites other than the respiratory system, and 3.7% had an underlying cause of death which was not cancer. For this latter group it is possible that cancer was mentioned on the death certificate as a contributing cause of death. This analysis is possible but was not carried out.



### 3.2.3 Availability and Completeness of the Data (Tables 5 and 6)

One measure of quality and usefulness of the data files is the frequency of valid information for specific data items. This measure is crude because "valid" as defined here means valid according to computerized edit checks and does not preclude that imputation of missing information or errors in definition or classification of the data item render the information invalid.

Subject to the above caveats the measure may be useful in showing if and where there are improvements in reporting of data items over the years, and whether or not particular analyses of the data are feasible. For example, information on date of birth is important for purely statistical (age-specific) analysis of the data as well as for medical follow-up analyses which depend on good identifying information.

On the cancer incidence file, a complete birthdate (i.e., day, month and year) is on average present on only 68% of the records in the period 1969-1978. If the two time periods, 1969-1973 and 1974-1978 are considered separately some improvement in the more recent period becomes evident.

On cancer mortality records for the same time period a complete birthdate is present in over 95% of cases. However, at least part of this high rate is due to the fact that the mortality system imputes a date of birth from age and date of death when the birthdate is not reported. In 1976 the imputation rate was 11.5% [5]. No such imputations are carried out in the cancer incidence system<sup>6</sup>.

Small area analyses of cancer occurrence require complete and detailed residence information. Cancer mortality data are much more useful for these purposes because census division (county) of residence codes are present on 99.8% of records and census subdivision (city, town, village) codes are present on 96.2% of records. In contrast, on the cancer incidence file, census division codes are present on 89.6% of records and census subdivision

---

<sup>6</sup> Imputations may be useful for statistical purposes but are actually mis-leading in medical follow-up studies unless it is made clear that the information is based on an imputation.

codes on only 25.2% of records. On the incidence file, there is improvement in the reporting of census subdivision information in the second time period.

### 3.2.4 Multiple Registrations on the Cancer Incidence System (Table 7)

Comparability of cancer incidence data is affected if there are differences in the reporting of multiple primary cancers in the same individual and in inadvertent duplicate registration.

The rules for reporting of multiple primary cancers are difficult to interpret, so some provincial differences in their application are expected.

Inadvertent duplicate registration of the same cancer incident may occur if a provincial registry cannot determine if the same case has been registered previously (perhaps because identifying information is inadequate) or if the same incident is reported by two different registries<sup>7</sup>. The search for multiple primary cancers was restricted to multiple entries for the same cancer site (at the 3-digit level of the ICDA code).

No attempt was made to separate duplicate registrations from multiple primaries, although it can be speculated that the majority of cases reported by two separate registries may be duplicates whereas the cases reported by the same registry are more likely to be valid multiple primaries, particularly those that differ in the 4th digit of the diagnosis code.

Using very strict matching criteria and excluding skin cancers (other than melanoma of the skin), 1.7% (6113) of records on the 1969-1978 cancer incidence file were identified as multiple entries.

By province, this rate varied from 0.5% to 1.9%. Only 0.4% of multiple records were reported by two different registries. There was agreement down to the 4th digit level of the diagnosis code in 88.4% of cases.

---

<sup>7</sup> The national Cancer incidence System does not carry out routine checks on such duplication.



By cancer site, if only sites with more than fifty records are considered, the rate of multiple primaries varied from 1.0% for cancer of the stomach and pancreas to 3.6% for breast cancer. The high rate for breast cancer is not surprising since the current rules for reporting of multiple primary cancers require separate reports for cancers in both sides of (most) bilateral organs.

On the whole it is felt that while there is some inconsistency arising from multiple primary and duplicate reporting, this is very small compared with that arising from undercoverage.

#### 4. SUMMARY

The techniques described in this paper have been successful at identifying differing levels of quality of cancer incidence and mortality data. It has been found that the mortality-incidence ratios, in particular, can be used to assess coverage errors, which are one of the major concerns of a high quality cancer incidence system. The data quality for those who are registered on the incidence system is sufficiently high that it is possible to assess the quality of the cause of death classification on the mortality system through a micro-data match. In fact a computerized micro-data match could be used to evaluate the undercoverage because the number of cancer deaths without previous registration on the NCIR system could be ascertained.

#### ACKNOWLEDGEMENTS

The authors are grateful to J. Gorman, D. Lawrence, K. McClean, S. Moore and P. Walsh for their assistance in preparing the data for this paper. We are also grateful to J. Silins for his comments. The referee's comments are also very much appreciated.

REFERENCES

- [1] Waterhouse, J., Muir, C., et al., Editors, CANCER INCIDENCE IN FIVE CONTINENTS, VOLUME III, WHO, International Agency for Research on Cancer, Lyon, 1976, p.3.
- [2] WHO HANDBOOK FOR STANDARDIZED CANCER REGISTRIES, World Health Organization, Geneva, 1976.
- [3] MANUAL FOR CANCER RECORDS OFFICERS, National Cancer Institute of Canada.
- [4] Op. cit., in reference 1 , pp. 45-51
- [5] Nagnur, D.N., Currie, S.G., Heath, B., QUALITY ASSESSMENT OF VITAL STATISTICS (A pilot study), Statistics Canada, Health Division, Ottawa, 1981.
- [6] King, H.S., Wigle, D.T., Hill, G.B., Silins, J., MORTALITY TRENDS FOR CANCERS OF THE CORPUS UTERI AND CERVIX UTERI, Alberta 1969-1978, CMAJ.
- [7] Doll, R., Peto, R., THE CAUSES OF CANCER: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today, JNCI, Vol. 66, No. 6, June 1981.
- [8] Percy, C., COMPARISON OF THE CODING OF DEATH CERTIFICATES RELATED TO CANCER IN SEVEN COUNTRIES, Public Health Reports, Vol. 93, No. 4, July-August, 1978.
- [9] Feigl, P., Breslow, N.E., Laszlo, J., Priore, R.L., Taylor, W.F., U.S. CENTRALIZED CANCER PATIENT DATA SYSTEM FOR UNIFORM COMMUNICATION AMONG CANCER CENTRES, JNCI, Vol. 67, No. 5, November, 1981, p. 1019.

Table 1

Cancer Mortality - Incidence Ratios

Deaths in Period (Mortality) as Percentage of New Cases Registered (Incidence)  
Canada (excluding Ontario) and Provinces with the Highest and Lowest Ratios.

Males		Leading Sites by Age Group and Sex					
Cancer Site	Age	1969 - 1973			1974 - 1978		
		Canada	Highest Ratio	Lowest Ratio	Canada	Highest Ratio	Lowest Ratio
Lung (162)	Total	96	110	83	95	101	80
	0-24	27	37*	-	70	-	100*
	25-44	89	96	83	85	75	82
	45-64	94	106	81	88	86	78
	65+	100	118	83	101	118	82
Prostate (185)	Total	44	55	35	39	54	33
	0-24	71*	133	-	71*	-	100*
	25-44	39	54	-	22	-	-
	45-64	27	32	33	24	30	20
	65+	48	62	35	43	60	36
Colon (153)	Total	76	101	54	67	86	56
	0-24	56	60	-	40	-	-
	25-44	61	76	40	53	87	45
	45-64	66	86	41	58	61	56
	65+	84	115	61	73	100	58
Bladder (188)	Total	35	42	26	28	35	23
	0-24	20	25*	33*	3	-	-
	25-44	7	9	10	6	-	-
	45-64	23	30	16	18	22	13
	65+	44	53	32	36	45	29
Stomach (151)	Total	104	124	82	90	112	80
	0-24	29*	-	-	86*	-	-
	25-44	82	102	54	74	40	77
	45-64	93	102	79	80	89	80
	65+	112	144	86	96	130	80
All Cancers (140-209) excluding skin (173)	Total	69	83	56	64	77	56
	0-24	58	64	45	48	49	56
	25-44	51	62	39	45	55	37
	45-64	65	78	53	60	65	50
	65+	74	94	59	69	90	60

- Either mortality, incidence or both are zero.

\* Ratios based on fewer than 10 cases for both mortality and incidence.

Table 2

Cancer Mortality - Incidence Ratios

Deaths in Period (Mortality) as Percentage of New Cases Registered (Incidence)  
Canada (excluding Ontario) and Provinces with the Highest and Lowest Ratios.

Females		Leading Sites by Age Group and Sex					
Cancer Site	Age	1969 - 1973			1974 - 1978		
		Canada	Highest Ratio	Lowest Ratio	Canada	Highest Ratio	Lowest Ratio
Breast (174)	Total	40	47	31	37	46	31
	0-24	14	13	-	6	50*	-
	25-44	28	33	26	23	30	17
	45-64	37	44	31	35	43	29
	65+	50	62	35	46	55	38
Colon (153)	Total	73	94	50	67	80	50
	0-24	15	20	-	27	33*	-
	25-44	48	49	41	41	64	34
	45-64	63	77	43	55	57	42
	65+	83	114	55	74	95	55
Uterus (182)	Total	27	38	17	20	26	15
	0-24	33	40*	20*	-	-	-
	25-44	13	19	7	11	15	9
	45-64	16	25	9	12	15	10
	65+	49	64	33	35	43	26
Cervix Uteri (180)	Total	39	45	29	32	48	27
	0-24	12	20*	17*	5	-	4
	25-44	21	17	12	15	25	10
	45-64	39	54	32	34	49	33
	65+	66	64	51	53	65	50
Ovary (183)	Total	69	79	54	69	78	63
	0-24	20	33*	-	28	100*	-
	25-44	43	41	26	30	38	20
	45-64	68	74	52	64	72	55
	65+	87	105	77	95	98	106
All Cancers (140-209) excluding skin (173)	Total	57	67	46	53	64	45
	0-24	46	49	43	37	44	28
	25-44	33	39	27	27	35	22
	45-64	48	56	38	44	51	39
	65+	75	94	58	69	85	58

- Either mortality, incidence or both are zero.

\* Ratios based on fewer than 10 cases for both mortality and incidence.

Table 3

Match of Mortality with Incidence Records

Disease Classification on the Cancer Incidence File for a Sample of Lung Cancer Deaths  
(Percentage Distribution)

Canada (excluding Ontario) and  
Provinces with the Highest and Lowest Rates of Lung Cancer Incidents among the Matches.

ICDA-8	CLASSIFICATION ON THE INCIDENCE FILE	CANADA	HIGHEST	LOWEST
	1. CANCER OF THE RESPIRATORY SYSTEM	95.2	100.0	83.7
162.1	(a) LUNG; Primary	92.5	100.0	74.4
160-163	(b) OTHER RESPIRATORY SYSTEM; Primary	1.0		4.7
197.0-197.3	(c) RESPIRATORY SYSTEM; Secondary	1.7		4.7
	2. OTHER CANCERS	4.8		16.3
174	(a) BREAST	0.7		2.3
200-209	(b) LYMPHATIC PHEMATOPOIETIC SYSTEM Primary	0.7		0.0
196	Secondary	1.0		4.6
	(c) OTHER SPECIFIED PRIMARY SITE	1.9		9.3
195, 199	(d) ILL DEFINED OR UNDEFINED SITE	0.5		0.0
	TOTAL	100.0	100.0	100.0

SAMPLE SIZE FOR MATCHES	415	54	43
MATCH SUCCESS RATE (%)	82.3	96.4	76.8



Table 4

Match of Incidence with Mortality Records

Cause of Death Classification for a Sample of Lung Cancer Cases from the Cancer Incidence File  
(Percentage Distribution)

Canada (excluding Ontario) and  
Provinces with the Highest and Lowest Rates of Lung Cancer Deaths among the Matches.

ICDA-8	CLASSIFICATION ON THE MORTALITY FILE	CANADA	HIGHEST	LOWEST
	1. CANCER OF THE RESPIRATORY SYSTEM	92.0	100.0	85.3
162.1	(a) LUNG; Primary	91.4	100.0	82.9
160-163	(b) OTHER RESPIRATORY SYSTEM; Primary	0.3		0.0
197.0-197.3	(c) RESPIRATORY SYSTEM; Secondary	0.3		2.4
	2. OTHER CANCERS	4.3		9.8
174	(a) BREAST	0.3		0.0
	(b) LYMPHATIC PHEMATOPOIETIC SYSTEM			
200-209	Primary	0.3		0.0
196	Secondary	-		0.0
	(c) OTHER SPECIFIED PRIMARY SITE	3.1		9.8
195, 199	(d) ILL DEFINED OR UNDEFINED SITE	0.6		0.0
	3. NOT CANCER	3.7		4.9
	TOTAL	100.0	100.0	100.0

SAMPLE SIZE FOR MATCHES	350	45	41
MATCH SUCCESS RATE (%)	69.4	80.4	73.2

Table 5

**Cancer Incidence  
Availability and Completeness of Data Items**

Canada (excluding Ontario) and  
Provinces with the Highest and Lowest Percentages of Data Completeness

DATA ITEM	YEAR OF DIAGNOSIS	CANADA	HIGHEST PERCENT	LOWEST PERCENT
<b>DATE OF BIRTH</b>				
Day	1969 - 1973	63.7	99.8	0.0
	1974 - 1978	71.6	99.8	4.0
	1969 - 1978	68.0	99.3	2.2
Month	1969 - 1973	65.6	98.8	0.0
	1974 - 1978	73.3	99.8	4.1
	1969 - 1978	69.8	99.4	2.2
Year	1969 - 1973	92.9	100.0	15.9
	1974 - 1978	96.3	100.0	21.9
	1969 - 1978	94.7	100.0	19.1
Complete Birthdate	1969 - 1973	63.6	98.7	0.0
	1974 - 1978	71.6	99.8	4.0
	1969 - 1978	68.0	99.3	2.2
<b>AGE</b>	1969 - 1973	99.4	100.0	98.6
	1974 - 1978	100.0	100.0	100.0
	1969 - 1978	99.7	100.0	99.4
<b>BIRTHPLACE</b> (Country or Province)	1969 - 1973	15.2	19.8	0.0
	1974 - 1978	24.4	71.1	0.0
	1969 - 1978	20.2	46.0	0.0
<b>RESIDENCE</b>	1969 - 1973	89.6	100.0	4.1
Census Division	1974 - 1978	89.6	100.0	82.8
	1969 - 1978	89.6	100.0	49.0
Census Subdivision	1969 - 1973	16.6	43.2	0.0
	1974 - 1978	32.3	76.4	0.0
	1969 - 1978	25.2	61.6	0.0

Table 6

**Cancer Mortality**  
**Availability and Completeness of Data Items**

Canada (excluding Ontario) and  
Provinces with the Highest and Lowest Percentages of Data Completeness

DATA ITEM	YEAR OF DEATH	CANADA	HIGHEST PERCENT	LOWEST PERCENT
<b>DATE OF BIRTH</b>				
Day	1969 - 1973	96.6	99.4	0.0
	1974 - 1978	97.8	99.7	43.9
	1969 - 1978	97.2	99.6	23.0
Month	1969 - 1973	97.0	99.7	0.0
	1974 - 1978	98.1	99.9	44.2
	1969 - 1978	97.6	99.8	23.2
Year	1969 - 1973	100.0	100.0	99.2
	1974 - 1978	99.9	100.0	99.6
	1969 - 1978	99.9	100.0	99.4
Complete Birthdate	1969 - 1973	96.6	99.4	0.0
	1974 - 1978	97.8	99.7	43.9
	1969 - 1978	97.2	99.6	23.0
AGE	1969 - 1973	100.0	100.0	100.0
	1974 - 1978	99.9	100.0	99.6
	1969 - 1978	100.0	100.0	99.8
BIRTH PLACE (Country or Province)	1969 - 1973	98.3	100.0	98.9
	1974 - 1978	52.2	99.9	0.0
	1969 - 1978	73.9	99.9	46.2
RESIDENCE Census Division	1969 - 1973	99.8	100.0	99.6
	1974 - 1978	99.7	100.0	99.8
	1969 - 1978	99.8	100.0	99.7
Census Subdivision	1969 - 1973	92.2	99.9	51.9
	1974 - 1978	99.7	99.5	99.4
	1969 - 1978	96.2	99.7	77.1

Table 7

Cancer Incidence  
1969 - 1978  
Multiple Primaries Within Each Site  
(Canada excluding Ontario)

ICDA Cancer Site	Multiple Primaries			
	Number	Percent of Incidence	Percent Same Registry	Percent With Same 4 <sup>th</sup> Digit ICDA Code
<b>Total All Sites (except skin, 173)</b>	<b>6,113</b>	<b>1.7</b>	<b>76.8</b>	<b>88.4</b>
140 Lip	87	1.4	90.8	83.9
141 Tongue	30	1.6	63.3	63.3
142 Salivary Gland	8	0.6	12.5	75.0
143 Gum	8	1.5	100.0	62.5
144 Floor of Mouth	18	1.8	94.4	N.A.
145 Mouth, Other and Unspecified	16	1.4	75.0	62.5
146 Oropharynx	14	1.0	92.9	78.6
147 Nasopharynx	8	1.1	62.5	N.A.
148 Hypopharynx	3	0.5	100.0	33.3
149 Pharynx, Unspecified	3	1.3	66.7	N.A.
150 Oesophagus	37	1.1	75.7	N.A.
151 Stomach	178	1.0	71.9	74.2
152 Small Intestine	12	1.2	91.7	91.7
153 Lge. Intestine Excl. Rectum	623	1.9	82.3	44.0
154 Rectum	256	1.4	77.0	74.6
155 Liver	11	0.6	63.6	90.9
156 Gall Bladder	22	0.7	77.3	77.3
157 Pancreas	96	1.0	70.8	54.2
158 Peritoneum	6	0.7	100.0	100.0
159 Unspec. Digestive Organs	1	0.3	100.0	N.A.
160 Nose, Etc.	5	0.7	80.0	80.0
161 Larynx	94	2.0	77.7	57.4
162 Trachea, Bronchus, Lung	795	1.9	75.0	99.5
163 Resp. Organs, Other & NOS	7	0.7	85.7	85.7
170 Bone	28	2.0	78.6	82.1
171 Connective Tissue	32	1.3	81.3	75.0
172 Melanoma of Skin	61	1.4	78.7	54.1
174 Breast	1,791	3.6	84.4	N.A.

Table 7 (concl'd)

Cancer Incidence  
1969 - 1978  
Multiple Primaries Within Each Site  
(Canada excluding Ontario)

ICDA Cancer Site	Multiple Primaries			
	Number	Percent of Incidence	Percent Same Registry	Percent With Same 4 <sup>th</sup> Digit ICDA Code
180 Cervix Uteri	133	1.4	60.9	N.A.
181 Chorionepithelioma	1	1.0	0.0	N.A.
182 Other, of Uterus	149	1.1	73.8	91.3
183 Ovary, Etc.	126	1.5	80.2	97.6
184 F. Genital Organs, Other	31	1.5	71.0	83.9
185 Prostate	464	1.6	69.2	N.A.
186 Testis	27	1.5	44.4	N.A.
187 M. Genital Organs, Other	9	1.4	88.9	100.0
188 Bladder	274	1.6	77.7	N.A.
189 Urinary Org., Other & NOS	119	1.5	72.3	79.8
190 Eye	13	1.1	46.2	N.A.
191 Brain	93	1.6	46.2	N.A.
192 Other Nervous System	6	0.4	33.3	66.7
193 Thyroid Gland	49	1.5	55.1	N.A.
194 Other Endocrine Glands	5	0.6	80.0	100.0
195 Ill - Defined Sites	4	0.6	100.0	100.0
196 Sec. & Unspec. Lymph Nodes	5	0.3	100.0	100.0
197 Sec., Resp. & Digestive	9	0.3	88.9	77.8
198 Other Secondary	1	0.1	100.0	100.0
199 Without Spec. of Site	12	0.3	75.0	83.3
200 Lymphosarcoma, Etc.	68	1.1	60.3	97.1
201 Hodgkin's Disease	85	2.2	55.3	N.A.
202 Other of Lymphoid Tissue	22	0.8	90.9	72.7
203 Multiple Myeloma	45	1.3	64.4	N.A.
204 Lymphatic Leukaemia	64	1.5	62.5	84.4
205 Myeloid Leukaemia	32	1.0	53.1	90.6
206 Monocytic Leukaemia	4	1.0	75.0	100.0
207 Other & Unspec. Leukaemia	11	0.7	81.8	90.9
208 Polycythemia Vera	1	0.2	100.0	N.A.
209 Myelofibrosis	1	0.4	100.0	N.A.



## ESTIMATING MONTHLY GROSS FLOWS IN LABOUR FORCE PARTICIPATION<sup>1</sup>

Stephen E. Fienberg and Elizabeth A. Stasny<sup>2</sup>

The Canadian Labour Force Survey is a household survey conducted each month for the purpose of producing point-in-time estimates of the number of persons employed, unemployed and not in the labor force. The survey has a rotating panel design in which all individuals in a sampled household location are interviewed each month, for six consecutive months. In the past, little use has been made of this longitudinal structure, although considerable interest has been expressed in the month-to-month gross flows (transitions) amongst the labour force status categories. In this paper we discuss methods being considered by Statistics Canada for the production of gross flow estimates, but from a model-based perspective.

### 1. INTRODUCTION

The Canadian Labour Force Survey is a monthly household survey used to produce cross-sectional or point-in-time estimates of labour force participation. This survey, however, like the Current Population Survey in the United States and many other large scale sample surveys, is designed using a panel structure so that the subjects are interviewed a number of times before being dropped from the sample. Although the survey is used mainly to obtain cross-sectional estimates, it has long been recognized that information from the repeated interviewing of subjects provides an additional longitudinal data base that could be exploited to give estimates of change over time for a very small additional cost (see, for example Kalachek, 1979, and Fienberg and Tanur, 1983).

---

<sup>1</sup> This research was supported in part by a contract with Statistics Canada. The authors wish to thank Murray Lawes, Larry Swain and Richard Veevers for their help in understanding the Labour Force Survey Methodology and the resulting data, as well as the Editor and a referee for helpful comments and suggestions.

<sup>2</sup> Stephen E. Fienberg and Elizabeth A. Stasny, Carnegie-Mellon University, Pittsburgh, Pennsylvania, PA. 15213.

Some attempts have been made to use the longitudinal data obtained from panel surveys. For example, the longitudinal data from the Current Population Survey has been used since 1948 to produce tables showing gross movements of individuals between labor force categories from one month to the next. Although these tables are produced each month, they have not been published since 1952 because of statistical problems. Smith and Vanski (1979) discuss the production of gross change data using the longitudinal structure of the Current Population Survey.

Recently, Statistics Canada has initiated an investigation of possible uses of the longitudinal data available as a by-product from the Canadian Labour Force Survey. They, too, would like to find a method for producing reliable estimates of gross movements between labour force categories. In this paper, we discuss the methods being considered by Statistics Canada for the production of such gross change data.

In Section 2, we give a brief description of the coverage and design of the Labour Force Survey, and we describe the structure of the resulting data. Then in Section 3 we outline the proposed method for gross flow estimation developed by Statistics Canada, which depends on the use of sample-based weights, adjustment for inflows to and outflows from the population of interest, consistency adjustments, and bias correction for misclassification error. By developing some simple models for the gross flow process, we explore in Section 4 the implications of Statistics Canada's proposed method. Finally, in Section 5 we describe some work on handling non-response in gross flow estimation.

## 2. DESCRIPTION OF THE LABOUR FORCE SURVEY

### 2.1 Survey Coverage

Approximately 56,000 households, chosen from the ten provinces of Canada, are included in the Labour Force Survey sample each month. Questionnaires are

completed for all civilian, non-institutionalized members of sampled households who are 15 years of age and older. The survey questions primarily relate to the subjects' work related activities during the reference week which is the week prior to the survey week and usually contains the fifteenth day of the month. Responses to the survey questions are used to classify subjects as employed, unemployed, or not in the labour force. For a discussion on classification of labour force status, see Guide to Labour Force Survey Data, Statistics Canada, (1979).

## 2.2 Survey Design

The labour Force Survey was designed to enable estimation of levels and rates of employment and unemployment for each of the ten provinces separately. Thus, except for the constraint on the total sample size, each province is sampled independently.

Economic regions (ER's), areas of similar economic structure, form the primary strata within provinces. ER's are divided into self-representing units (SRU's) and non-self-representing units (NSRU's). SRU's are large urban centers and NSRU's are generally composed of a small urban center and a rural area. Sampling is carried out separately in SRU's and NSRU's.

SRU's are sampled using a stratified, two-stage sampling design. NSRU's are sampled using a stratified multi-staged sampling scheme. In addition to the SR and NSR areas, some sample units are selected from an apartment frame and a special area frame. The final sampling units for the Labour Force Survey are households. A detailed description of the sampling plan for the survey can be found in Methodology of the Canadian Labour Force Survey 1976, Statistics Canada (1977).

Households selected for the Labour Force Survey are included in the survey for six consecutive months and are then dropped from the sample. For example, households rotated into the survey in January are interviewed for six consecutive months, and then dropped from the sample after the June interview. Each

group of households that rotated into and out of the sample together makes up a panel. In any one month, individuals from six different panels are included in the Labour Force Survey sample.

### 2.3 Sampled-Based Weights

The cross-sectional data, information for a given month from all subjects in the six panels interviewed in that month, is used to produce monthly estimates of labour force participation. The monthly estimates are weighted averages of values for each person in the sample. A weighted average is used because each sampled person is thought of as "representing" a number of people in the population of interest. The weight assigned to an individual's record corresponds to the number of persons in the population that the person in the sample represents.

Let  $W_{t,i}$  be the weight assigned to surveyed individual  $i$  in the month  $t$ . If individual  $i$  is classified as outside the population of interest in month  $t$ , then  $W_{t,i} = 0$ . Otherwise, the assigned weights are determined by the probability of selecting the cluster, the probability of selecting the household within the cluster, nonresponse within the month, rural/urban factors, sub-sampling adjustment for fast-growing areas, and ratio adjustments for province/age/sex factors.

An individual's assigned weight can change from month to month because of replacement of 1/6 of the sample each month, non-response, and, to a lesser extent, because of changes in the size of the population of interest. Thus, for any one individual,  $i$ , it may be the case that  $W_{t-1,i} = W_{t,i}$ .

### 2.4 Longitudinal Structure and Gross Flow Estimation

Although the main purpose for the Labour Force Survey is to produce point-in-time estimates of labour force participation, the panel structure of the survey design results in a longitudinal data base, with approximately 5/6 of the



household locations sampled in any one month being in the sample for the following month. Naturally, fewer than 5/6 of the surveyed individuals or families are the same in two consecutive months due to non-response and moving. However, Statistics Canada is interested in the possibility of using information from those individuals who do respond in two consecutive months to produce estimates of gross flows among labour force categories.

Estimates of gross flows are useful for answering questions such as a) How much of the increase in unemployment is due to persons losing jobs and how much is due to persons formerly not in the labour force starting to look for jobs? or b) How many unemployed persons become discouraged and leave the labour force?

We discuss the problem of estimating gross flows among labour force categories in the next two sections.

### 3. STATISTICS CANADA'S PROPOSED METHOD FOR GROSS FLOW ESTIMATION

In this section we describe the multi-stage estimation procedure for gross flows developed by Statistics Canada (e.g. see Macredie and Veevers, 1977; Wong, 1983). Our description of the procedure includes various interpretations of the impact of individual stages.

#### 3.1 Data Needed to Estimate Gross Flows

Statistics Canada has proposed estimating gross flows using a 4x4 matrix as shown below:



Labour Force Status in Month t

		E	U	N	O
Labour Force Status in Month t-1	E	$x_{EE}$	$x_{EU}$	$x_{EN}$	$x_{EO}$
	U	$x_{UE}$	$x_{UU}$	$x_{UN}$	$x_{UO}$
	N	$x_{NE}$	$x_{NU}$	$x_{NN}$	$x_{NO}$
	O	$x_{OE}$	$x_{OU}$	$x_{ON}$	$x_{OO}$

where E = employed

U = unemployed

N = not in the labour force

O = outside the population of interest, and

$x_{ij}$  = estimated number with labour force status i in month t-1 and status j in month t.

The final monthly Labour Force Survey files from two consecutive months can be used to obtain the data for estimating the 4x4 matrix of gross flows. In order to use these data to produce gross flow estimates, Statistics Canada must match individual records from the two consecutive monthly files using the unique identification numbers assigned to sampled individuals for the duration of their time in the study.

An individual appearing on the data file for one month may be missing from the file for the other month due to rotation into or out of the sample or because the person moved, was not at home, or refused to respond. The sample weights described in Section 2.4 include an adjustment for non-response within each month. When dealing with gross flows, we also need to consider the month-to-month non-response. Statistics Canada proposes to reweight records for individuals who responded in both months t-1 and t to compensate for this additional non-response.

After the reweighting is completed, Statistics Canada will have a single data file that includes information for all persons who were respondents in two consecutive months. That file will contain geographic and demographic information for each individual as well as the individual's labour force status and assigned weights for both month  $t-1$  and month  $t$ .

### 3.2 Differences in Weights

As we noted in Section 2.3, an individual's sample-based weight can change from month to month because of the rotation replacement structure, non-response, and changes in the size of the population of interest. Even when the adjustment factor for non-response is computed on the basis of month-to-month data, for any individual  $i$ , it may still be the case that  $W_{t-1,i} \neq W_{t,i}$ . Some method is needed for handling this difference in weights if data are to be used for estimating gross flows.

Statistics Canada proposes to resolve this dilemma by assuming that differences in the two weights occur only as a result of inflows to and outflows from the population of interest. Thus, differences in weights are added to the appropriate cell in either the last row or last column of the gross flow matrix. This procedure depends heavily on the interpretation of the weights suggested in Section 2.3, namely that sample individual  $i$  represents  $W_{t,i}$  persons in the population in month  $t$ .

As an illustration of the procedure, suppose an individual is classified as employed in both months  $t-1$  and  $t$  but  $W_{t-1,i} = 300$  and  $W_{t,i} = 305$ . The minimum weight, 300, is added to the EE cell of the gross flow table. The difference,  $W_{t,i} - W_{t-1,i}$ , of 5 is added to the OE cell since those 5 would be thought of as having been outside the population of interest during the month  $t-1$  and then having moved in the population as employed persons for month  $t$ .

If, on the other hand, the individual is employed in both months but  $W_{t-1,i} = 305$  and  $W_{t,i} = 300$ , then 300 is again added to the EE cell but the excess weight of 5 is added to the EO cell. Here, the difference between weights represents 5 persons who were employed in month  $t-1$  and then moved outside the population of interest in month  $t$ .

An individual who is classified as outside the population of interest in month  $t-1$  and is then, say, employed in month  $t$  will have  $W_{t-1,i} = 0$ . If  $W_{t,i} = 300$ , then 300 is added to the OE cell. Individuals classified as outside the population of interest in month  $t$  are treated similarly with  $W_{t-1,i}$ , being added to the appropriate cell in the last column of the gross flow matrix.

Because persons outside the population of interest are assigned a weight of zero, a person who is classified as such in both months  $t-1$  and  $t$  would have  $W_{t-1,i} = 0$  and  $W_{t,i} = 0$ . Therefore,  $X_{00}$ , the entry in the 00 cell of the gross flow matrix, must always be zero.

### 3.3 Adjustment of the Inflow and Outflow Cells

Adding differences in weights to the inflow and outflow cells of the gross flow matrix provides a method for handling the changes in sample-based weights from one month to the next and gives estimates of inflows to and outflows from the population of interest. Independent estimates of inflows and outflows, available from Census data, suggest that this method overestimates the actual amount of movement into and out of the population of interest. Thus, Statistics Canada plans to adjust the  $X_{OE}$ ,  $X_{OU}$ ,  $X_{ON}$ ,  $X_{EO}$ ,  $X_{UO}$ , and  $X_{NO}$  entries in the gross flow matrix. These cells will be proportionally adjusted so that total inflows and outflows shown in the gross flow matrix equal the Census estimates of inflows and outflows respectively.

Let  $I$  be the independent census estimate of inflows to the population of interest and  $F$  be the census estimate of outflows from the population. Call the

sum of estimated inflows  $X_{0+} = X_{OE} + X_{OU} + X_{ON}$  and the sum of estimated outflows  $X_{+0} = X_{EO} + X_{UO} + X_{NO}$ . The proportionally adjusted inflows are:

$$Y_{0j} = X_{0j}I/X_{0+} \quad \text{for } j = E, U, N. \quad (1)$$

The proportionally adjusted outflows are

$$Y_{j0} = X_{j0}F/X_{+0} \quad \text{for } j = E, U, N. \quad (2)$$

### 3.4 Consistency of Gross Flow Estimates With Monthly Totals

Statistics Canada would like their gross flow estimates to be consistent with the published monthly estimates of labour force participation totals. Thus, the row totals for the gross flow matrix must be the month  $t-1$  estimates of labour force participation and the column totals must be the month  $t$  cross-sectional estimates. The marginal totals of the gross flow matrix constructed as described above are not consistent with the monthly labour force totals.

Statistics Canada plans to use the method of iterative proportional scaling, originally proposed by Deming and Stephan (1940), and described in detail by Bishop, Fienberg, and Holland (1975), to adjust the gross flow matrix to agree with the monthly labour force totals. When used to adjust the gross flow matrix, iterative proportional scaling alternatively 1) constrains the rows of the matrix to sum to the month  $t-1$  estimates and then 2) constrains the columns to sum to the month  $t$  estimates. Steps 1) and 2) are repeated until the entries in the matrix do not change from one step to the next.

Testing at Statistics Canada has shown that cell changes resulting from the application of iterative proportional scaling were both absolutely and relatively small and fell roughly within the bounds of sampling variability associated with the cells. This suggests that the consistency adjustment does not seriously distort the gross flow estimates.



### 3.5 Bias Correction for Misclassification Error

Statistics Canada proposed method for estimating gross flows also includes a step correcting for misclassification bias. This is the bias that results from the incorrect assignment of an individual's labour force status. A technique developed by Fred Wong (1983) at Statistics Canada uses reinterview data to correct for the misclassification bias.

## 4. IMPLICATIONS OF STATISTICS CANADA'S PROPOSED METHOD

### 4.1 Modeling Gross Flows

Each step of Statistics Canada's proposed method described in the previous section is a logical attempt to correct problems that arise concerning the production of good estimates of gross flows. It is not clear, however, what effect the various adjustments have on the final estimated gross flow matrix. In order to better understand Statistics Canada's proposal to treat differences in weights as being due to inflows to and outflows from the population of interest, in this section we develop a model for the gross flow process. Our discussion centers on the quantities in the inflow and outflow cells of the estimated gross flow matrix since the problems with Statistics Canada's proposed method seem to occur in those cells. Because the design of the Canadian Labour Force Survey is quite complex, we begin with a set of simplifying assumptions. In the following we assume that

1. a single stage stratified sample is chosen,
2. there is no response error, and
3. non-response occurs only because random individuals move between strata or because of rotation into or out of the sample.

### 4.2 Allocation of Net Population Changes to Inflow and Outflow Cells

Suppose that the population of interest is divided into  $S$  strata indexed by  $s = 1, 2, \dots, S$ . Let



$N_k^s$  = population size in strata  $s$  in month  $k$ .

Each month, a simple random sample is chosen from each stratum for the survey and sampled individuals are interviewed for six consecutive months before being dropped from the survey. Our goal is to estimate gross flows from month  $t-1$  to month  $t$ .

For the purpose of estimating gross flows, only individuals who are interviewed in both month  $t-1$  and  $t$  will be used. This excludes individuals who rotate into or out of the sample and persons who move between strata. Let

$r_{t-1,t}^s$  = number of sampled individuals from stratum  $s$  who were interviewed in both months  $t-1$  and  $t$ .

Each of the  $r_{t-1,t}^s$  respondents in stratum  $s$  is assigned the following weights in months  $t-1$  and  $t$  respectively for the purpose of gross flow estimation:

$$w_{t-1}^s = N_{t-1}^s / r_{t-1,t}^s \text{ and } w_t^s = N_t^s / r_{t-1,t}^s. \quad (3)$$

As long as movements between strata and selection of panels are "random processes," these weights represent the inverse of the probability that an individual within a stratum is interviewed in both months  $t-1$  and  $t$ . Since all individuals within a stratum have the same weight in any given month, aggregates for each stratum may be used. Therefore, we let

$n_{ij}^s$  = number of sampled individuals from stratum  $s$  classified as having labour force status  $i$  in month  $t-1$  and status  $j$  in month  $t$  for  $i, j = E, U, N, O$ .

The methodology proposed by Statistics Canada requires that the minimum of the months  $t-1$  and  $t$  weights for each individual be added to the appropriate cell in the gross flow matrix. The difference is added to the appropriate inflow

cell if the month  $t$  weight is greater than the weight in month  $t-1$  and to the appropriate outflow cell otherwise. Thus, for example, the stratum  $s$  entry in the EE (employed to employed) cell of the gross flow matrix is:

$$\begin{aligned} \min(W_{t-1}^s, W_t^s) n_{EE}^s &= \min [(N_{t-1}^s / r_{t-1,t}^s), (N_t^s / r_{t-1,t}^s)] n_{EE}^s \\ &= \min (N_{t-1}^s, N_t^s) n_{EE}^s / r_{t-1,t}^s \\ &= \min (N_{t-1}^s, N_t^s) f_{EE}^s \end{aligned} \quad (4)$$

where  $f_{EE}^s$  = fraction of all individuals from stratum  $s$ , interviewed in both months  $t-1$  and  $t$ , who were employed in both months.

The contribution from stratum  $s$  to the OE cell for the matrix from individuals employed in month  $t$  is:

$$\begin{aligned} \max(0, W_t^s - W_{t-1}^s) n_{EE}^s &= \max[0, (N_t^s - N_{t-1}^s / r_{t-1,t}^s)] n_{EE}^s \\ &= \max(0, N_t^s - N_{t-1}^s) n_{EE}^s / r_{t-1,t}^s. \end{aligned} \quad (5)$$

Differences from individuals falling in the UE and NE cells will also contribute to the OE cell. Thus, the total contribution to the OE cell from stratum  $s$  is:

$$\begin{aligned} \max(0, N_t^s - N_{t-1}^s) \{ (n_{EE}^s / r_{t-1,t}^s) + (n_{UE}^s / r_{t-1,t}^s) + (n_{NE}^s / r_{t-1,t}^s) \} \\ &= \max(0, N_t^s - N_{t-1}^s) n_{+E}^s / r_{t-1,t}^s \\ &= \max(0, N_t^s - N_{t-1}^s) f_{+E}^s \end{aligned} \quad (6)$$

where  $f_{+E}^s$  = fraction of all individuals from stratum  $s$ , interviewed in both months  $t-1$  and  $t$ , who were employed in month  $t$ .

We obtain total for all cells in the gross flow matrix in a similar manner. The resulting gross flow matrix is as follows:

Gross Flow Matrix - Month  $t-1$  to Month  $t$

		Month $t$				
		E	U	N	O	
Month $t-1$	E	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{EE}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{EU}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{EN}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{E+}^s$	$\sum_{s=1}^S N_{t-1}^s f_{E+}^s$
	U	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{UE}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{UU}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{UN}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{U+}^s$	$\sum_{s=1}^S N_{t-1}^s f_{U+}^s$
	N	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{NE}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{NU}^s$	$\sum_{s=1}^S \min(N_{t-1}^s, N_t^s) f_{NN}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{N+}^s$	$\sum_{s=1}^S N_{t-1}^s f_{N+}^s$
	O	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+E}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+U}^s$	$\sum_{s=1}^S \max(0, N_{t-1}^s - N_t^s) f_{+N}^s$	0	
		$\sum_{s=1}^S N_t^s f_{+E}^s$	$\sum_{s=1}^S N_t^s f_{+U}^s$	$\sum_{s=1}^S N_t^s f_{+N}^s$		

Notice that each term in the summations for the nine in-population cells (the cells showing gross flows between employed, unemployed, and not in the labour force) is the product of the net size of the strata and the observed fraction of subjects who had the various labour force classifications in months  $t-1$  and  $t$ . The out-of-population to employed cell of the gross flow matrix contains a sum of terms from each strata that grew from month  $t-1$  to month  $t$ . Each term is the product of the net increase in size for the strata and the fraction of the subjects in the strata who reported being employed in month  $t$ . The out-of-population to unemployed and not in the population of interest cells

contain the sums of similar terms except that the net increase in size for each strata is multiplied by the fraction of subjects in the strata who were unemployed or not in the labour force in month  $t$  respectively. In other words, the net increase in size for each strata is proportionally allocated to the three inflow cells of the gross flow matrix based on the observed fractions of employed, unemployed, and not in the labour force in month  $t$ . Similarly, the net decrease in size for each strata that shrank between months  $t-1$  and  $t$  is proportionally allocated to the outflow cells of the matrix based on the observed fractions of employed, unemployed, and not in the labour force in month  $t-1$ .

In this model we have assumed that the only way for counts to appear in the inflow and outflow cells is as a result of differences in weights. In practice, a small number of individuals who move in and out of the population-of-interest show up in the sample and their assigned weights are added to the appropriate inflow and outflow cells. The effect of such individuals on the estimates is very small.

The fractions  $f_{+E}^S$ ,  $f_{+U}^S$ ,  $f_{+N}^S$ ,  $f_{E+}^S$ ,  $f_{U+}^S$ , and  $f_{N+}^S$  are estimated using individuals who appear in the sample in both months. Almost all individuals classified, for example, as OE could not be respondents in both months because they are not sampled by design or because they are movers. Thus, these people who could not have been respondents in both months are represented by people who did respond in both months. To the extent that these groups differ, the proportional allocation of net increases and decreases in strata size may result in biased estimates in the inflow and outflow of the gross flow matrix.

#### 4.3 Effects of Movements Between Strata

The weights used for the purpose of gross flow estimation, as shown in expression (3), are determined by the number of respondents in both months  $t-1$  and  $t$ , a quantity that remains constant for the two months, and by the stratum population. The population of a stratum changes if a) individuals enter from

outside the population of interest, such as when persons reach their 15th birthday or leave the full-time military, b) individuals move outside the population of interest, as when persons enter the military or an institution, or c) individuals in the population of interest move between strata. This subsection describes the effects of such changes in population size on the quantities in the gross flow matrix.

As in the preceding subsection, we suppose that the population of interest is divided into  $S$  strata. Again, individuals are sampled at random from each stratum every month, interviewed for six consecutive months, and then dropped from the sample. Let  $r_{t-1,t}^s$  be, as before the number of individuals from stratum  $s$  who are interviewed in both months  $t-1$  and  $t$ .

Next we suppose that there are  $N_{t-1}^s$  individuals in stratum  $s$  in month  $t-1$ . Let movements into and out of strata between months  $t-1$  and  $t$  be denoted by

$m_{u,v}$  = number of individuals who move from  $u$  to  $v$ ,  $u \neq v$ , between interviews for months  $t-1$  and  $t$  where  $u$  and  $v$  may take on the values.

$s$  = stratum  $s$  for  $s = 1, 2, \dots, S$  and

$0$  = outside the population of interest.

Using this notation, the population in stratum  $s$  in month  $t$  is

$$N_t^s = N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u}). \quad (7)$$

The weights assigned to individuals in stratum  $s$  in months  $t-1$  and  $t$  respectively are

$$w_{t-1}^s = N_{t-1}^s / r_{t-1,t}^s \text{ and } w_t^s = N_t^s / r_{t-1,t}^s. \quad (8)$$



Since our focus in this section is on movement into and out of the population of interest, it is not necessary for us to divide those in the population of interest into employed, unemployed, and not in the labour force. Thus, the gross flow matrix used here is a  $2 \times 2$  matrix formed by collapsing the first three rows and columns of the  $4 \times 4$  gross flow matrix used in the preceding subsection.

The entry for stratum  $s$  in the in-population to in-population cell is

$$\begin{aligned} \min(W_{t-1}^s, W_t^s) r_{t-1,t}^s &= \min \{ N_{t-1}^s / r_{t-1,t}^s, [N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})] / r_{t-1,t}^s \} r_{t-1,t}^s \\ &= \min [N_{t-1}^s, N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})] \\ &= N_{t-1}^s + \min [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]. \end{aligned} \quad (9)$$

The entry for stratum  $s$  in the out-of-population to in-population, or inflow, cell is

$$\begin{aligned} \max(0, W_t^s - W_{t-1}^s) r_{t-1,t}^s &= \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u}) / r_{t-1,t}^s] r_{t-1,t}^s \\ &= \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]. \end{aligned} \quad (10)$$

The entry for the in-population to out-of-population, or outflow, cell is found similarly. Thus, the  $2 \times 2$  gross flow matrix is as follows:

		Month t		
		In-Population	Out-of-Population	
Month t-1	In population	$\sum_{s=1}^S \{N_{t-1}^s + \min[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]\}$	$\sum_{s=1}^S \max[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]$	$\sum_{s=1}^S N_{t-1}^s$
	Out-of-population	$\sum_{s=1}^S \max[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]$	0	

$$\sum_{s=1}^S \{N_{t-1}^s + \sum_{u \neq s} (m_{u,s} - m_{s,u})\}$$

Let us consider the quantity in the inflow cell of this gross flow matrix. This cell should contain the net increase in population from outside the population of interest,  $m_{0,s} - m_{s,0}$ , for each stratum that gained members from outside the population. What the cell does contain is  $\sum_{u \neq s} (m_{u,s} - m_{s,u})$  for each stratum  $s$  that grew as a result of movements between strata and from outside the population of interest. The summation,  $\sum_{u \neq s} (m_{u,s} - m_{s,u})$ , does include the quantity  $m_{0,s} - m_{s,0}$  but it may also contain other terms.

For example, suppose the population is made up of three strata called A, B, and C. If strata A and B grew from month  $t-1$  to month  $t$  and stratum C lost members, then the inflow cell contains

$$\begin{aligned} \sum_{u \neq A} (m_{u,A} - m_{A,u}) + \sum_{u \neq B} (m_{u,B} - m_{B,u}) &= m_{0,A} - m_{A,0} + m_{B,A} - m_{A,B} + m_{C,A} - m_{A,C} \\ &+ m_{0,B} - m_{B,0} + m_{A,B} - m_{B,A} + m_{C,B} - m_{B,C} \end{aligned}$$

$$= m_{0,A} - m_{A,0} + m_{0,B} - m_{B,0} + m_{C,A} - m_{A,C} + m_{C,B} - m_{B,C} \quad (11)$$

Note that the movements between strata A and B cancel out but the terms showing the movement between strata A and C and strata B and C remain in the summation.

In general, the inflow cell contain extra terms of the form  $m_{v,u} - m_{u,v}$  for each stratum  $v$  that loses population while stratum  $u$  gains population. Similarly, the outflow cell contains extra terms of the form  $m_{x,y} - m_{y,x}$  for each stratum  $y$  that grows while stratum  $x$  loses population.

In the inflow cell, the quantity  $\sum_{u \neq s} (m_{u,s} - m_{s,u})$  for each strata  $s$  that gains population from month  $t-1$  to  $t$  will be positive, although each individual term in the summation need not be positive. If

$$\sum_{u \neq s} (m_{u,s} - m_{s,u}) > m_{0,s} - m_{s,0} \quad (12)$$

then the contribution to stratum  $s$  is more than the inflow to stratum  $s$  from outside the population of interest. This excess comes from terms of the form  $m_{u,v} - m_{v,u}$  as described above. That is, the overestimate is due to movements between strata within the population. A similar result holds for the in-population to out-of-population cell of the matrix.

Statistics Canada staff report that the method they proposed for handling differences in weights from month to month does appear to give overestimates in the inflow and outflow cells of the gross flow matrix. Although they are based on simplifying assumptions, the results here give a possible explanation for the overestimation, i.e. the overestimation may be due to movements within the population of interest.

Finally we note that, in the  $2 \times 2$  gross flow matrix shown above, the in-population to in-population cell must contain an underestimate equal to the overestimate in the outflow cell. Whatever the amount of underestimation, it is spread over the nine in-population to in-population cells in the  $4 \times 4$  gross flow matrix. Moreover, the size of the overestimation is small in comparison

to the total size of the nine in-population cells.

#### 4.4 Comments on the Proposed Gross Flow Estimation Method

The results described in the preceding two subsections illustrate problems with the proposed method of handling month-to-month differences in weights for the purpose of gross flow estimation. These results do not come as a surprise to Statistics Canada. Because of their experience with Labour Force Survey methods and data, they realized that the movements in individuals within the population might explain some of the overestimation in the inflow and outflow cells of the gross flow matrix. The results obtained by modelling the process reinforce their beliefs and make it clear just how the movements of individuals effects the estimates. In addition, the modelling brought to light a problem about which Statistics Canada had not been aware: the compensating underestimation spread over the nine in-population to in-population cells of the gross flow matrix.

In section 4.2, we saw that the net increases in strata are allocated to the inflow cells while the net decreases are allocated to the outflow cells according to the fractions of observed individuals classified as employed, unemployed, and not in the labour force in month  $t$  and month  $t-1$ , respectively. The estimation of inflows and outflows in this manner is valid only if individuals who move in and out of the population of interest are a random sample of individuals and, hence, "the same" as individuals who remain within the population of interest. Sampled individuals who are classified as outside the population of interest appear in the sample by accident rather than by design; the Labour Force Survey is not designed to estimate numbers of persons outside the population of interest. If we want to obtain reasonable estimates for the inflow and outflow cells of the matrix, it may be necessary to include individuals outside the population of interest in the Labour Force Survey sample or to use a special, supplementary sample.

In section 4.3, we saw that the overestimates in the inflow and outflow cells

could be a result of movements of individuals between a strata whose population grew and a strata whose population shrank. The fact that it was movements between strata that caused the problem is a result of the simplifying assumptions which we made. We assumed that the final sample was randomly chosen from within each strata. Hence, the weights assigned to individuals sampled from a single strata were equal. If, instead, we assumed that the strata had been divided into clusters and random samples of individuals had been chosen from within the clusters, then all individuals sampled from a single cluster would have been assigned the same weight and the overestimate would come as a result of movements between clusters.

To correct for the overestimate, and corresponding underestimate, directly in the case where final samples are chosen at random from within strata, we would need estimates of the number of movers between each pair of strata where one strata grew and the other strata lost population. If the final samples are chosen at random from within clusters, similar estimates would be required for each pair of clusters. This is a considerable amount of information. A further complication is that, in practice, the ratio adjustments applied to the weights make it possible for individuals within one household to have unequal weights.

As was suggested earlier, if individuals outside the population of interest were included in the sample, we could obtain estimates of movement into and out of the population of interest directly. One other possibility that should be considered is to discard the monthly weights for the purpose of gross flow estimation and derive a longitudinal weight for each individual in the Labour Force Survey sample in either of the two months.

As statisticians, we are quite comfortable with estimates of gross flows that do not have the published monthly labour force participation totals as marginal totals; however, we realize the problems that might arise if gross flow estimates, that were not consistent with the monthly totals, were published. Nevertheless, it should not be assumed automatically that the monthly estimates are correct and that the problem lies solely in the gross flow estimates.



As we noted in section 3.5, the gross flow matrix is adjusted to correct for misclassification errors. The monthly estimates, however, are not corrected for misclassification bias. Thus, when iterative proportional fitting is used to adjust the gross flow matrix to agree with the monthly totals, the matrix is being altered to be consistent with biased values. We feel that it would be more appropriate to address the problem of misclassification of labour force status in the monthly data where it occurs rather than just in the estimates of gross flows.

## 5. NON RESPONSE AND GROW FLOW ESTIMATION

Statistics Canada's proposed method for gross flow estimation compensates for non-response by adjusting the sample-based weights of respondents. This method of handling non-response is appropriate if the missing data are missing at random (e.g., see Rubin, 1976). In order to explore the assumption of random non-response, we used a longitudinal file for a single panel to produce the data in Table 1. This table shows the unweighted percentages of individuals reporting being employed or unemployed in zero to six months according to the number of months in which they responded to the survey.

Consider the probabilities underlying the observed percentages shown in part (a) of Table 1. Let

$\pi_i$  = probability that an individual is employed in  $i$  out of 6 months for  $i = 0, 1, \dots, 6$ .

Under the assumption that non-response occurs at random, the probabilities corresponding to the first column of that table can be written as

$P(\text{observing } 0 \text{ months employed out of } 6-k \text{ months responding})$

$$= \frac{\sum_{j=1}^k \pi_j}{\binom{6}{j}}, \quad \text{for } k = 0, 1, \dots, 5. \quad (13)$$

Notice that these probabilities increase from the first row of the column to the last row.

In a similar manner, it can be shown that, if data are missing at random, then the underlying probabilities must increase from the top to bottom of each column in both tables. The first column of each table deviates from this pattern quite noticeably. In both cases, the observed percentages decrease through the first four rows of the table and then increase in the last two rows. It does not seem likely that sampling variability alone could be responsible for such a pattern in both tables. Thus, it appears as if there is some evidence that non-response does not occur at random.

Of course, the above analysis is based on just a single panel of Labour Force Survey data. However, in a larger study using data from 1980 and 1981, Paul and Lawes (1982) also found evidence of a relationship between employment status and non-response. Therefore, there is a need to consider methods for gross flow estimation that do not require the assumption that non-response occurs at random.

Statistics Canada's proposed method for gross flow estimation only makes use of the information from individuals who responded in both of the months. There is also information available from those individuals who responded in just one of the two months. Stasny (1983) presents a method for month-to-month gross flow estimation that makes use of the information available from individuals who are respondents in only one of the two months and that can be used when non-response is related to time or employment status. For this method, we take the observed gross flow data to be the end result of a two-stage process. In the first stage of the process, which we do not get to observe, individuals are allocated to the sixteen cells of the gross flow matrix according to a single multinomial sampling scheme. Then, in the second stage, each individual may lose either the month  $t-1$  or month  $t$  labour force classification with some probability. The probability of losing a month's classification can be modeled to depend on the month, or labour force status, or both. Maximum likelihood estimates for the parameters of the multinomial

distribution of the first stage and the probabilities of losing a month's classification are obtained using iterative methods.

When these models were fit to Labour Force Survey data from a single panel, Stasny (1983) found that the model where the probability of losing a month's classification depends on labour force status provides a reasonable fit to the data for all gross flow matrices with the exception of the months 1-2 matrix. For the data from month 1 to month 2, the probability of losing a month's classification appears to depend on the month. This may be due to the fact that there is higher non-response in the first month a panel is in the survey. We believe that it would be worthwhile to fit this type of model to additional data from the Labour Force Survey to see if similar results are obtained over other panels.

Clearly, the problem of obtaining good estimates of gross flows from the Labour Force Survey is not a simple one. The survey is designed to give data for the production of monthly estimates of labour force participation, not estimates of gross flows. A survey designed specifically for the purpose of estimating gross flows among labour force categories would certainly be different from the current Labour Force Survey. Thus, the longitudinal data from the survey is not ideal for gross flow estimation. The data, however, are available and, if they can be used to give reasonable estimates of gross flows, then additional, useful information is produced for a relatively small cost.

#### REFERENCES

- [1] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. Discrete Multivariate Analysis: Theory and Practice. The MIT Press, 1975.
- [2] Deming, W.E. and Stephan, F.F. On A Least Squares Adjustment Of a Sampled Frequency Table When the Expected Marginal Totals Are Known. Annals of Mathematical Statistics, 1940, 11, 427-444.

- [3] Fienberg, S.E. and Tanur, J.M. The Design and Analysis of Longitudinal Surveys: Controversies and Issues of Cost and Continuity. Technical Report 289, Department of Statistics, Carnegie-Mellon University, May 1983.
  
- [4] Kalachek, E. Longitudinal Surveys and Labor Market Analysis. Data Collection, Processing and Presentations: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 160-189.
  
- [5] Macredie, I.D. and Veevers, R. Discussion on Smith, R.E. and Vanski, J.E. Gross Change Data: The Neglected Data Base. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 153-158.
  
- [6] Paul, E.C. and Lawes, M. Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey. Survey Methodology, 1982, 8, 48-85.
  
- [7] Rubin, D.B. Inference and Missing Data. Biometrika, 1976, 63, 581-592.
  
- [8] Smith, R.E. and Vanski, J.E. Gross Change Data: The Neglected Data Base. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 132-150.
  
- [9] Stasny, E.A. Estimating Gross Flows in Labor Force Participation Using Information From Individuals With Incomplete Classifications. Technical Report 272, Department of Statistics, Carnegie-Mellon University, Jan. 1983.

- [10] Statistics Canada, Guide to Labour Force Survey Data. Catalogue 71-528 Occasional, July 1979.
- [11] Statistics Canada, Methodology of the Canadian Labour Force Survey 1976. Catalogue 71-526 Occasional, Oct. 1977.
- [12] Wong, F.A. Technique To Correct The Response Bias In the 4x4 Labour Force Gross Flow Matrix, Technical Reprot, Statistics Canada, 1983.



Table 1

		Months Employed						
		0	1	2	3	4	5	6
a)								
	6	52.23	3.22	1.93	2.21	2.45	3.55	34.41
	5	51.48	3.04	2.17	3.36	4.22	35.73	
Months of	4	49.26	4.15	3.16	4.83	38.60		
Data Present	3	46.31	6.18	5.62	41.89			
	2	51.40	8.32	40.28				
	1	52.87	47.13					
		Months Unemployed						
		0	1	2	3	4	5	6
b)								
	6	92.43	3.63	1.65	0.93	0.57	0.43	0.35
	5	91.23	4.46	2.08	1.13	0.72	0.38	
Months of	4	89.28	5.76	2.17	1.43	1.36		
Data Present	3	89.00	6.42	2.81	1.77			
	2	91.33	6.01	2.66				
	1	91.43	8.57					

## REDESIGN OF THE NIAGARA TENDER FRUIT OBJECTIVE YIELD SURVEY

J. Kovar<sup>1</sup>

The peach, sour cherry and the grape objective yield surveys have been carried out annually in the Niagara Peninsula since 1964 in order to forecast the magnitude of change in marketable fruit production from the previous year. Timeliness of the estimates is essential in order to enable the Ontario Tender Fruit Growers Marketing Board (OTFGMB) and the Ontario Grape Growers Marketing Board (OGGMB) to establish the marketing strategies well ahead of the harvest. This paper summarizes the major changes due to the second redesign initiated in 1982. In particular, the sample design, data collection operation and modifications of the estimation procedures are elaborated upon.

### 1. INTRODUCTION

The decision to switch from a list frame to an area frame survey was made in the first redesign in 1974 primarily due to the lack of an adequate list of commercial growers in the Niagara Peninsula. However, in 1981 the Ontario Ministry of Agriculture and Food (OMAF) has conducted a Tree Fruit Census and a Grape Vine Census. The availability of the census data makes it possible to redesign the survey for the second time in order to reflect the changes that the industry has undergone in the last eight years. Based on discussions with OMAF, it was decided that the census lists of growers are complete and accurate and that they contain sufficient information to form the sampling frame for the Tender Fruit Surveys. As a result, the peach, sour cherry and grape surveys will be conducted employing three independent samples selected from the 1981 OMAF census lists.

The object of all three surveys is to forecast the total amount of fruit actually sold (as fresh fruit or to processors). These forecast are made by

---

<sup>1</sup> J. Kovar, Business Survey Methods Division, Statistics Canada. This work was done while the author was in the Institutional and Agriculture Survey Methods Division, Statistics Canada.

estimating a ratio of the number of pieces of marketable fruit in the current year to the corresponding total for the previous year and applying this ratio to the previous year's figure of actual amount sold reported as a tonnage by the Ontario Fruit and Vegetable Statistics Committee. Thus an assumption of high correlation of fruit weight and fruit count must be made. Secondly due to the time lag between the surveys and the harvest, it has to be assumed that any loss of fruit between these two times is consistent from year to year.

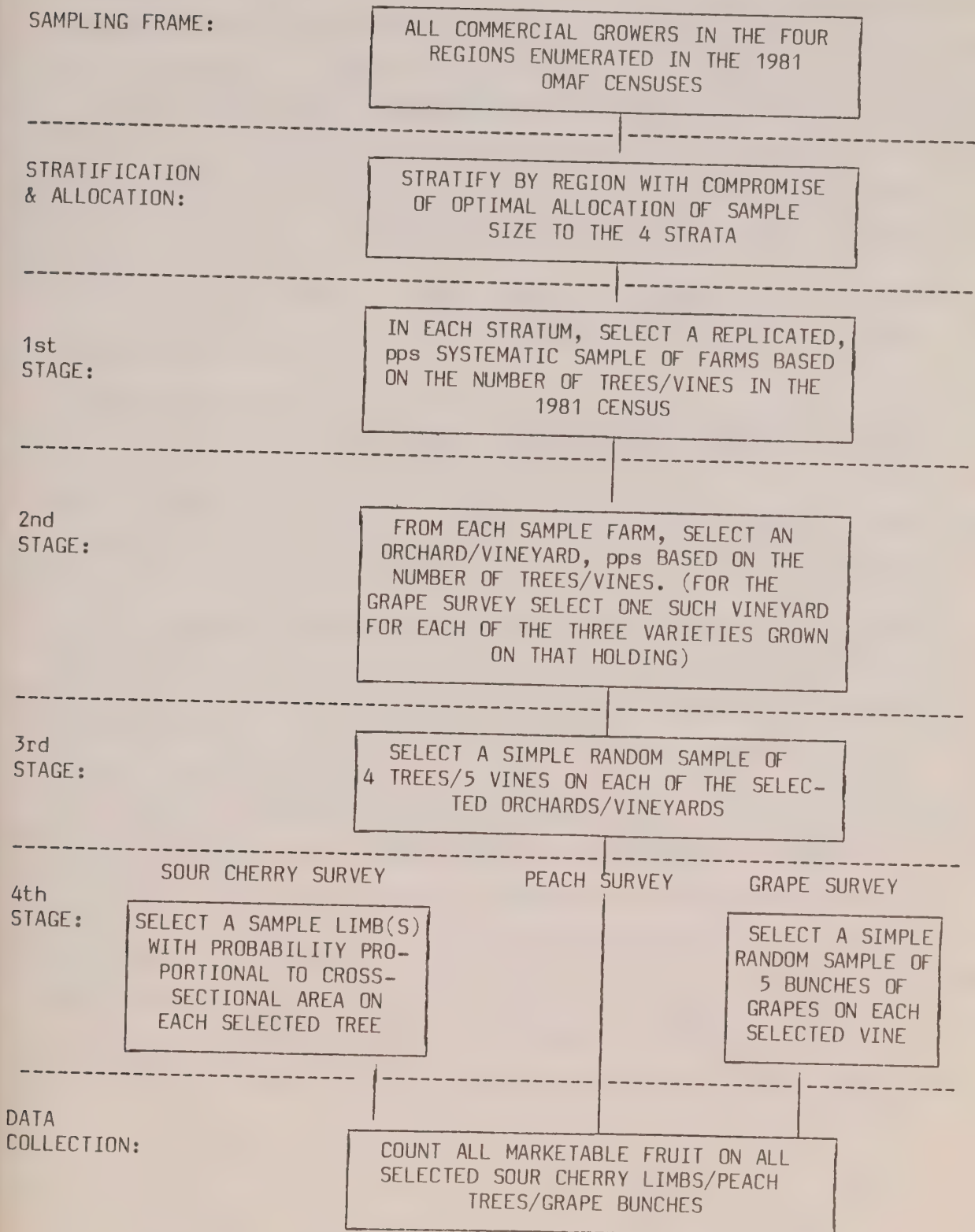
## 2. OVERVIEW OF THE SAMPLE DESIGN

The samples of the three objective yield surveys were selected independently according to a multistage, stratified (by geographical region), replicated, pps (farms and orchards/vineyards were selected with probability proportional to size), nearly self-weighting (all trees/vines have an approximately equal probability of selection) sample design. Figure 1 provides a visual summary of the sampling strategy. Note that due to the fact that the weight variables are collected at various points in time, the design is not exactly self-weighting.

### 2.1 Target Population, Sampling Frames and Total Sample Size

The target population for the three objective yield surveys comprises all commercial growers of the respective fruit in the Niagara Peninsula. Commercial growers for the three surveys were defined by OMAF as operators of those holdings which reported more than 200 peach trees, 200 sour cherry trees or 5000 grape vines respectively in the 1981 Tree Fruit or Grape Vine Censuses. Using the above definition, a separate frame was created for each of the three surveys. The lists for the peach, the sour cherry and the grape surveys contain 423, 145 and 552 commercial growers respectively. The total sample size (number of orchards/vineyards to be enumerated) for each survey was determined by OMAF's budget constraints to be in the neighbourhood of 60 for the peach survey, 55 for the sour cherry survey and 155 for the grape survey. Since for the grape survey all available varieties of interest are to be sampled on a selected farm, the final sample size for the grape survey is unknown. However, based on the 1981 Grape Vine Census, it is estimated that 62 farms will generate a sample of approximately 155 vineyards.

FIGURE 1: Sample Design for the Tender Fruit  
Objective Yield Surveys



## 2.2 Stratification and Sample Size Allocation to Regions

The Niagara Peninsula was divided into four regions for which separate estimates are required. These were defined as follows (based on the 1976 Census boundaries):

Region 1: Town of Grimsby in the Niagara Regional Municipality and township of Saltfleet in the Regional Municipality of Hamilton-Wentworth. (Township 8 of county 29 and township 4 of county 17).

Region 2: City of St. Catharines and the Town of Lincoln in the Niagara Regional Municipality. (Townships 5 and 9 of county 29).

Region 3: Town of Pelham and the Town of Thorold in the Niagara Regional Municipality. (Townships 11 and 12 of county 29).

Region 4: City of Niagara Falls and the Town of Niagara-on-the-Lake in the Niagara Regional Municipality. (Townships 3 and 10 of county 29).

Due to the increasing demand of crop production estimates by geographic area, an independent sample of farms was drawn in each of the four regions. An attempt was made to allocate the resources (i.e. number of farms sampled) optimally between regions. However, due to the unusually small population size in some regions (see Table 1) a compromise between proportional allocation, optimal allocation and a rule of "minimum of 2 farms per region per replicate" was made. The latter rule was deemed appropriate in order to diminish the possibility of complete nonresponse in a given replicate (as could be the case if only one farm per replicate was selected). The number of trees/vines in each farm was used as a measure-of-size variable for the purposes of allocation as well as for pps selection in the first and second stages. Previous results [6] indicate that other proxy variables (such as area under cultivation) are likely to be no more efficient than the tree-count variable.



TABLE 1: Population and Sample Sizes for the Tender Fruit Surveys of Commercial Growers by Region

	REGION 1		REGION 2		REGION 3		REGION 4		TOTAL	
	POPL'N SAMPLE		POPL'N SAMPLE		POPL'N SAMPLE		POPL'N SAMPLE		POPL'N SAMPLE	
PEACH	15	4	198	22	15	4	195	30	423	60
SOUR CHERRY	20	4	55	20	30	20	40	12	145	56
GRAPE	67	4	275	32	46	4	164	22	552	62

### 2.3 First Stage Design

Within each region, for each survey, two independent replicates of farms were selected systematically (in order to obtainm a representative sample) with probability proportional to the total number of trees/vines on the holding as of the 1981 Censuses. The total sample sizes for the two replicates are displayed in Table 1 by region. Since the two replicates are selected independently and since large farms are more likely to be selected in the sample, it is to be expected that a certain amount of overlap between replicates will exist. In fact, some farms are so large, that not only are they guaranteed to be in the sample, but they can appear more than once in the same replicate [4]). Each such appearance is treated as a separate event and one orchard/vineyard is selected without replacement every time the farm is selected. The actual number of distinct farms in the sample is therefore decreased as indicated in Table 2.

TABLE 2: Total Number of Distinct Farms in the Sample by Region

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PEACH	3	22	3	27	55
SOUR CHERRY	4	17	15	10	46
GRAPE	4	30	4	20	58

## 2.4 Second Stage Design

From the second stage on, the sampling strategies involve some field operation. Once an initial contact with the farmer is made (in the spring of 1983) it is imperative that every effort be made to obtain the respondent's cooperation. It is at this time that the farmer will be requested to aid the enumerator in listing all orchards and establishing the current size (i.e. number of trees) of each for the peach and sour cherry surveys. For the grape survey a similar listing must be prepared for each of the three varieties of interest: Concord, DeChaunac and "Other". (Note that since some varieties are grown together, one vineyard can appear on several of the lists. However, its size for a particular variety listing would be measured by the number of vines of that variety only).

On each holding, for the peach and sour cherry surveys, one orchard will be selected with probability proportional to size. For the grape survey one vineyard will be selected independently for each of the three varieties actually grown on that holding, again with probability proportional to size. It cannot be overemphasized that these procedures must be followed faithfully in order not to jeopardize the validity of the estimates. Selection procedures should be monitored carefully to ensure there is no bias in the selection towards small orchards or single variety vineyards, which admittedly would be easier to enumerate.

To avoid an overlap of orchards/vineyards on farms that are selected in both replicates or more than once in the same replicate, all orchards/vineyards for a given holding are to be selected at the same time, using a pps systematic sampling method. The assignment to replicates is to be performed at random after this selection. (Note that for the grape survey, on farms which appear in both replicates, two vineyards of each variety grown are to be selected).

## 2.5 Third Stage Design

Once an orchard/vineyard is selected, its current count of producing trees/vines is determined and a simple random sample of four producing trees/five producing vines is selected without replacement. The trees/vines

are marked for future identification since the same units are enumerated from year to year. (In subsequent years, if a sampled tree/vine has been destroyed, pulled up or has died, a replacement tree/vine is selected and enumerated. However, it does not contribute to the estimate until its second year in the sample). Also, each year the producing tree/vine count of selected orchards/vineyards is reestablished in order that the industry's growth (decline) can be monitored. (Note that in the grape survey this implies that only vines of the particular variety sampled are to be counted in each vineyard).

## 2.6 Fourth Stage Design

This stage exists only for the sour cherry and grape surveys.

### 2.6.1 The Sour Cherry Survey

It is operationally impossible to count all sour cherries on a selected tree. To estimate the total marketable fruit count, a sample limb (or limbs) is selected with probability proportional to the cross-sectional area of the limb. A method of selecting a limb in this way is described by Jessen [3]. It consists of selecting a limb at the initial (or primary) branching point of the trunk with probability proportional to the cross-sectional area and following the selected limb to the next branching point. This is repeated until the cross-sectional area of a subsequently selected limb is within five to fifteen percent of the primary limbs cumulative cross-sectional area total. As it is not always possible to select one such limb, in some instances two limbs will have to be enumerated. The selected limb(s) on each sample tree are then marked for future identification since the same limbs are enumerated from year to year.

### 2.6.2 The Grape Survey

As for sour cherries, it is equally impossible to count all marketable berries on a sample vine. Thus to estimate this total, the number of bunches of grapes (i.e. those clusters containing more than five berries) is counted and 5 bunches are selected at random without replacement in order to be enumerated.

As with the other surveys, the vines are marked and are to be visited the following year.

### 3. DATA COLLECTION

The actual enumeration will be performed roughly four weeks before harvest each year. It is of great importance that the selected sample vines, trees and limbs as well as the orchards and vineyards be well identified in order to enable the enumerators to complete their job in the short time available. The enumerators will be required to count all marketable fruit (i.e. excluding culls which are immature or damaged fruit that will not be harvested) on the sample peach trees and the selected cherry limbs. The fruit on the entire peach tree is counted primarily due to the fact that the fruit tends to be distributed much more unevenly on a peach tree than on a sour cherry tree [2], precluding the possibility of merely enumerating sample limbs.

For the grape survey, all berries on the five selected bunches are to be counted, excluding culls. Since most bunches are very tightly packed, this will, in most cases, involve picking the fruit. For this reason and due to time constraints, it is impossible to enumerate the entire sample vine.

### 4. REPLACEMENTS

Even though every attempt will be made to return to the same trees, limbs or vines in the following years, there arise cases when this is impossible. (For example, branches are sawn off, trees or vines are pulled up or are otherwise destroyed). If a sour cherry tree limb was sawn off, an attempt will be made to select another limb on the same tree using the same procedures as before. In the event that this is not possible, then just as in the case of peach trees and grape vines, a new tree/vine will be selected at random in the same orchard/vineyard. In the case that the whole orchard/vineyard has been destroyed a new orchard/vineyard will be selected on the same holding using the same procedures as described in Section 2.4. In all these cases, the newly



selected sample limbs, trees, vines, orchards or vineyards will be enumerated, however they will will not contribute to the estimate until the following year's data is collected, as only matched observations are considered.

For those hopefully rare, cases where the farmer has ceased to grow the fruit of interest entirely or where the initial contact resulted in a refusal, a third "replicate" of much smaller size was selected without replacement for each of the surveys. The procedures for selecting the orchard/vineyard and the sample of trees, limbs and vines for each replacement farm are the same as those described above. The limbs, trees and vines will be enumerated every year but will contribute to the estimate only when it is necessary to rotate one of them into the sample. The sizes of the replacement sample are indicated in Table 3 by region.

Table 3: Sample Sizes of the Replacement Sample  
for the Tender Fruit Survey by Region

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PEACH	1	2	1	3	7
SOUR CHERRY	1	2	2	2	7
GRAPE	1	3	1	2	7

## 5. ESTIMATION FORMULAE

### 5.1 Estimates of Fruit Count per Tree/Vine

Denote by  $y_{\tau}$  the total number of marketable fruit on a tree (vine)  $\tau$ .



Then for the peach survey,  $y_{\tau}$  is estimated by  $\hat{y}_{\tau}$ , the total number of marketable peaches counted on a sample tree  $\tau$ . For the sour cherry survey,  $y_{\tau}$  is estimated by

$$\hat{y}_{\tau} = \hat{y}_{\tau\ell} / p_{\ell} \quad (5.1.1)$$

where  $\hat{y}_{\tau\ell}$  is the total number of marketable sour cherries counted on the sample limb(s)  $\ell$  of the selected sample tree  $\tau$ ;

and  $p_{\ell}$  is the probability of selecting the sample limb(s)  $\ell$ . Finally for the grape survey,  $y_{\tau}$  is estimated by

$$\hat{y}_{\tau} = \frac{N_{\tau}}{n_{\tau}} \sum_{\ell=1}^{n_{\tau}} \hat{y}_{\tau\ell} \quad (5.1.2)$$

where  $N_{\tau}$  is the total number of bunches of grapes on the sample vine  $\tau$ ;

$n_{\tau}$  is the number of bunches of grapes that were enumerated on the sample vine  $\tau$  (typically  $n_{\tau} = 5$ );

and  $\hat{y}_{\tau\ell}$  is the number of berries on a bunch  $\ell$  of the sample vine  $\tau$ .

## 5.2 Regional Estimates of Fruit Count by Replicate

Denote by  $\hat{Y}_{ar}$  the estimated total number of marketable fruit in replicate  $r$  of region (area)  $a$  in the current year. For the grape survey,

$$\hat{Y}_{ar} = \sum_{v=1}^3 \hat{Y}_{arv} \quad (5.2.1)$$

where  $\hat{y}_{arv}$  is the estimated total grape count of variety  $v$  in replicate  $r$  of region  $a$ .

For the purpose of uniformity of the following formulae, for the sour cherry and peach surveys  $\hat{y}_{ar}$  and  $\hat{y}_{arv}$  can be used interchangeably, since there is only one variety of sour cherries and peaches to be estimated. (In other words, the subscript  $v$  can be ignored for the sour cherry and peach surveys). Then  $\hat{y}_{arv}$  can be estimated by

$$\hat{y}_{arv} = \frac{1}{n_{arv}^*} \sum_{f=1}^{n_{arv}} \sum_{b=1}^{n_{arvf}} \frac{N_a^{81}}{N_{arf}^{81}} \times \frac{N_{arvf}^{83}}{N_{arvfb}^{83}} \times \frac{N_{arvfb}^C}{n_{arvfb}} \sum_{\tau=1}^{n_{arvfb}} \hat{y}_{arvfb\tau} \quad (5.2.2)$$

with  $\hat{y}_{arvfb\tau}$  = the current year's estimated total number of marketable fruit on the tree/vine  $\tau$  (variety  $v$ ) in orchard/vineyard  $b$ , on farm  $f$ , in replicate  $r$ , of region  $a$  (i.e.  $\hat{y}_{\tau}$ );

$n_{arvfb}$  = the number of trees/vines (of variety  $v$ ) sampled in the current year in orchard/vineyard  $b$ , of farm  $f$ , in replicate  $r$ , of region  $a$  (typically  $n_{arvfb} = 4$  for the sour cherry and peach surveys and 5 for the grape survey);

$n_{arvf}$  = the number of orchards/vineyards (sampled for variety  $v$  in the current year) on farm  $f$ , in replicate  $r$  of region  $a$  (typically  $n_{arvf} = 1$  except for duplicates, i.e. large farms selected more than once in the same replicate);

$n_{arv}$  = the total number of distinct farms on which variety  $v$  was sampled in the current year) in replicate  $r$ , of region  $a$ ;

$n_{arv}^*$  = the total number of orchards/vineyards (sampled for variety  $v$  in the current year) in replicate  $r$  of region  $a$

$$(i.e. \quad n_{arv}^* = \sum_{f=1}^n n_{arvf});$$

$N_{arvfb}^C$  = the total current count of producing trees/vines (of variety v) in orchard/vineyard b, on farm f, in replicate r, of region a;

$N_{arvfb}^{83}$  = the total count of trees/vines (of variety v), in orchard /vineyard b, on farm f, in replicate r of region a as of the 1982/1983 mapping operation;

$N_{arvf}^{83}$  = the total count of trees/vines (of variety v), on farm f, in replicate r, of region a as of the 1983 listing operation;

$N_{arf}^{81}$  = the 1981 Census count of total trees/vines (all varieties) on farm f replicate r, of region a (supplied with the sample listing)

and  $N_a^{81}$  = the 1981 Census count of all trees/vines in region a (as per Table 4)

Table 4: 1981 Census Counts of Trees/Vines ( $N_a^{81}$ )  
for the Tender Fruit Surveys by Region

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PEACH	13,094	389,157	10,271	411,697	824,219
SOUR					
CHERRY	9,496	50,888	55,449	32,536	148,369
GRAPE	1,142,067	5,660,008	946,509	3,975,202	11,729,786

### 5.3 Regional Estimates of Ratio of Change and their Precision

The ratio of change in production in region a from the previous year, denoted by  $R_a$ , is estimated by

$$\hat{R}_a = \hat{Y}_a / \hat{X}_a \quad (5.3.1)$$

where  $\hat{Y}_a$  = the estimated total marketable fruit count (of peaches, sour cherries, grapes or grapes of variety v) in region a in the current year is given by

$$\hat{Y}_a = \frac{1}{2} \sum_{r=1}^2 \hat{Y}_{ar} \quad (5.3.2)$$

in the case of peaches, sour cherries, and total grapes of all varieties, and by

$$\hat{Y}_a = \hat{Y}_{av} = \frac{1}{2} \sum_{r=1}^2 \hat{Y}_{arv} \quad (5.3.3)$$

in the case of grapes by variety, and where  $\hat{X}_a$ ,  $\hat{X}_{ar}$ ,  $\hat{X}_{arv}$  are the corresponding previous year's estimates.

(Note that the subscript v can now be dropped as all estimates are treated in the same manner, be it peach, sour cherry, total grape or grapes by variety estimates.)

Define the variances of  $\hat{Y}_a$  and  $\hat{X}_a$  and their covariance by

$$V(\hat{Y}_a) = (\hat{Y}_{a1} - \hat{Y}_{a2})^2 / 4 = D_{ya}^2 / 4 \quad (5.3.4)$$

$$V(\hat{X}_a) = (\hat{X}_{a1} - \hat{X}_{a2})^2 / 4 = D_{xa}^2 / 4 \quad (5.3.5)$$

$$\text{Cov}(\hat{Y}_a, \hat{X}_a) = (\hat{Y}_{a1} - \hat{Y}_{a2})(\hat{X}_{a1} - \hat{X}_{a2}) / 4 = D_{ya} D_{xa} / 4$$

where the numeric subscripts refer to the replicate number. Then the variance,  $V(\hat{R}_a)$ , of the ratio of change estimate,  $\hat{R}_a$ , can be estimated by [1].

$$\begin{aligned}\hat{V}(\hat{R}_a) &= \frac{1}{\hat{\chi}_a^2} \{V(\hat{Y}_a) - 2 \hat{R}_a \text{Cov}(\hat{Y}_a, \hat{X}_a) + \hat{R}_a^2 V(\hat{X}_a)\} \\ &= \left\{ \frac{D_{ya}}{S_{xa}} - \frac{S_{ya} D_{xa}}{S_{xa}^2} \right\}^2\end{aligned}\quad (5.3.7)$$

with

$$\begin{aligned}S_{ya} &= \hat{Y}_{a1} + \hat{Y}_{a2} \\ D_{ya} &= \hat{Y}_{a1} - \hat{Y}_{a2}, \text{ etc}\end{aligned}\quad (5.3.8)$$

The coefficient of variation of  $\hat{R}_a$  is then estimated by

$$C\hat{V}(\hat{R}_a) = \frac{\{\hat{V}(\hat{R}_a)\}^{\frac{1}{2}}}{\hat{R}_a} \times 100\% \quad (5.3.9)$$

#### 5.4 Regional Estimates of Total Fruit Production and their Precision

Denoting by  $\chi_a^T$  the previous year's actual yield (tonnage) in region a and by  $\hat{Y}_a^T$  the corresponding current year estimate, then  $\hat{Y}_a^T$  is given by

$$\hat{Y}_a^T = \chi_a^T \hat{R}_a \quad (5.4.1)$$

with its coefficient of variation estimated by

$$C\hat{V}(\hat{Y}_a^T) = \frac{\chi_a^T \{\hat{V}(\hat{R}_a)\}^{\frac{1}{2}}}{\chi_a^T \hat{R}_a} \times 100\% = C\hat{V}(\hat{R}_a) \quad (5.4.2)$$

#### 5.5 Estimates of Total Fruit Production and their Precision

Denote by  $\hat{Y}^T$  the estimated total fruit production over all four regions in the current year. Then  $\hat{Y}^T$  is given by,



$$\hat{Y}^T = \sum_{a=1}^4 \hat{Y}_a^T \quad (5.5.1)$$

with a coefficient of variation estimated by

$$c\hat{V}(\hat{Y}) = \frac{\left\{ \sum_{a=1}^4 (\hat{Y}_a^T)^2 V(\hat{R}_a) \right\}^{\frac{1}{2}}}{\hat{Y}^T} \times 100\% \quad (5.5.2)$$

## 6. SUMMARY

The first enumeration will take place in 1983, however, it will not be until the summer of 1984 that the first estimates from the redesigned survey will be produced. For this reason, it will be necessary to conduct both the old and the new surveys in 1983 so that estimates will be available for that year.

Even though the survey was designed to be self-weighting, it is only approximately so due to the time differences of the 1981 Censuses, the 1983 initial listing operation and the subsequent enumerations. The estimation formulae presented in the previous section take these time differences into account. However, due to the appealing simplicity of the self-weighting estimate, an investigation of its performance has been proposed once the data becomes available.

## REFERENCES

- [1] Cochran, W.G. (1963). Sampling Techniques, John Wiley and Sons, New York.
- [2] Davidson, G. (1977). Redesign of the sour cherry, peach and grape objective yield surveys in the Niaqara Peninsula, Survey Methodology 3, 38-61.
- [3] Jessen, R.J. (1955). Determining the fruit count on a tree by randomized branch sampling, Biometrics 11, 99-109.

- [4] Kish, L. (1965). Survey Sampling, John Wiley and Sons, New York.
- [5] Murthy, M.N. (1967). Sampling Theory and Methods, Statistical Publishing Society, Calcutta.
- [6] Singh, U. and Sukhatme, B.V. (1980). Sampling for estimating production of fruit crops, Sankhya C42, 17-30.

A TIMELY AND ACCURATE POTATO ACREAGE ESTIMATE FROM LANDSAT:  
RESULTS OF A DEMONSTRATION<sup>1</sup>

R.A. Ryerson, J.-L. Tambay, R.J. Brown

and

L.A. Murphy, B. McLaughlin<sup>2</sup>

This paper describes the procedures used and results of a joint Canada Centre for Remote Sensing (CCRS) and Statistics Canada project to provide a timely potato acreage estimate for New Brunswick, a major potato producing province in Canada. The project has demonstrated that satellite imagery combined with more traditional potato area estimation procedures can lower respondent burden, produce timely crop distribution maps and produce reliable estimates for subregions.

## 1. INTRODUCTION

Earlier satellite remote sensing work in St. John Valley, New Brunswick by the Canada Centre for Remote Sensing (CCRS) and the New Brunswick Department of Agriculture has proved that both an accurate and low cost estimate of potato crop area could be made using satellite data (Mosher et al. 1978; Ryerson et al. 1979; Ryerson et al. 1980). Interest in this and other CCRS work on rapeseed-canola (Brown et al. 1980) resulted in Statistics Canada initiating a real-time demonstration using data from Landsat satellite in the 1980 crop year. Statistics Canada, the Federal agency responsible for crop data collection, wanted to compare traditional and satellite-derived estimates of crop area in the same region. Potatoes were selected as the subject crop, and the St. John Valley was the region.

---

<sup>1</sup> Originally presented at the fifteenth International Symposium on Remote Sensing of Environment, Ann Harbor, MI, May 1981.

<sup>2</sup> R.A. Ryerson and J. Brown, Canada Centre for Remote Sensing (CCRS), E.M.R., J.-L. Tambay, Business Survey Methods Division (this work was done while the author was in the Institutional and Agriculture Survey Methods Division), Statistics Canada, L.A. Murphy, Agriculture Statistics Division, Statistics Canada, and B. McLaughlin, Agriculture Statistics Division, Regional office at Truro, Nova Scotia, Statistics Canada.

The particular benefits of satellite remote sensing which are of interest to Statistics Canada are improving accuracy of estimates obtained from their regular surveys, possibly at more local levels, lowering of respondent burden by reducing the number and/or size of questionnaires, and the possibility of providing maps of small areas containing speciality crops to better plan sampling methods.

Following a summary of the main results in the next section, the balance of this paper outlines the remote sensing methodology used in this project, and describes the existing Statistics Canada data collection system, the project region, the ground data sample and data collection, the analysis of remotely sensed data and the verification and analysis of results.

## 2. MAIN RESULTS

Data collected by satellite were used to produce estimates of the potato area in the St. John Valley region of New Brunswick. These estimates, expanded to the provincial level, were within two percentage points of the Statistics Canada published estimate of 52,000 acres. The published estimate was based on the results of three independent surveys in the province.

The analysis of satellite data was done in real-time (almost instantaneously) at CCRS, as much of the work could be initiated prior to data acquisition. The Agriculture Enumerative Survey (A.E.S.) area sample provided the ground data needed to calibrate the system, and was used to obtain ratio and regression estimates which corrected for biases in the satellite classification of potato fields. Although the demonstration was not carried out in a production environment, the final estimates could have been produced less than two weeks after the satellite pass over the test region.

Problems in the satellite classification were caused by the presence of clouds (satellite nonresponse), and by the confusion of potatoes with "similar appearing" features on the analysis system. The first problem caused loss of

data, and required some imputation. The second problem was partly resolved by adjustments to the classification, and by the use of ratio and regression estimators.

A comparison of interviewer - collected ground data with data collected through aerial photographs of sample fields showed that some fields were missed by the A.E.S. interviewers, as these were not required for A.E.S. purposes. This resulted in the aerial photography data being used instead of A.E.S. data for the 1980 satellite estimates. The 1981 A.E.S. enumeration procedure were changed to accommodate both A.E.S. and remote sensing requirements.

As a result of the success of this demonstration, the experiment was repeated in 1981. In addition, a similar experiment was undertaken that year to estimate the canola acreage in the Peace River District of Alberta and British Columbia.

### 3. REMOTE SENSING USING THE LANDSAT SATELLITE

Remote sensing is the sensing or measuring of the characteristics of an object from a distance, usually from an aircraft or satellite. When satellite data are used, complete coverage of large areas can be provided quickly at a relatively low cost. Possible areas of application include agriculture, forestry, land utilisation, ice formations and general map making.

The United States National Aeronautics and Space Administration Landsat satellites provided the satellite coverage for this, and an earlier experiment in New Brunswick. Each Landsat satellite orbits the earth 14 times a day in a sun-synchronous orbit (permitting coverage of the earth to be done at the same local sun time). Light reflected from the ground is recorded on four narrow bands of the spectrum using a Multispectral Scanner (MSS). The data are transmitted in Canada to one of two receiving stations in Prince Albert,



Saskatchewan and in Shoe Cove, Newfoundland. A point on earth is covered once every 18 days by a Landsat satellite (every nine days if two satellites are used).

The CCRS Image Analysis (CIAS) analyses the data, received on standard products such as Computer Compatible Tapes (CCT's) covering areas of 25,600 square kilometers. The smallest units for which image data are defined are called picture elements, or pixels. Each pixel carries its own spectral signature, a measurement of its reflectance on the four spectral bands. The spectral signature will depend on the features present in the pixel (roads, crops, etc.), each of which carries its own signature. Crop signature is a function of plant structure, type of soil background visible, crop maturity, height, and leaf density, among other factors.

To obtain estimates of crop areas, it is necessary to identify each pixel belonging to a crop of interest. Large known fields of the crop of interest are located to train the system to identify the crop's signature. All pixels are then classified as belonging to the crop or not, based on their spectral signatures.

Areas for specific regions are obtained by counting the number of pixels inside the regions that are classified as belonging to the crop. Additional training may be done to cover pixels "missed" in the initial classification, or to further separate confusion crops, that is, crops whose spectral signature closely resembles that of the crop of interest.

Accurate ground data are needed, first, to locate large training fields for the crop of interest and second, to correct for any biases in the satellite classification. These data can be obtained by trained ground enumerators, or by using airborne imagery, which is interpreted by image analysts.

#### 4. CURRENT STATISTICAL DATA COLLECTION SYSTEM

Historically, Statistics Canada has used data obtained from annual mail non-probability surveys as the primary input into its crop estimation system. While these surveys are relatively inexpensive and can be completed quickly, they are limited by varying response rates and possible non-representativeness of respondents. Probability enumerative surveys were introduced in the mid-1970's to overcome some of these problems. These involve enumeration of a random sample of farmers by personal interviews. In 1980, Statistics Canada's estimates of potato area in New Brunswick were based on the results of three surveys: the Mail Survey, the Objective Potato Yield Survey (O.P.Y.S.), and the Agriculture Enumerative Survey (A.E.S.).

The Mail Survey questionnaires are mailed out in early June to all farmers listed on a Farm Register maintained by the Agriculture Statistics Division. Replies are compiled on a county basis and county estimates are derived by linking annual changes in reported potato acreages to census potato acreages for the county. The county estimates are aggregated to give provincial estimates by late June.

The O.P.Y.S. is a specialized mail and enumerative survey designed to estimate potato area and yields in the potato growing region of New Brunswick. The Survey is conducted in mid-July on a random sample of potato farmers selected from the Farm Register and potato area estimates are generated by mid-August.

The A.E.S. is a multi-purpose enumerative survey designed to estimate crop, livestock and farm expense data at provincial levels. The A.E.S. is a multiple frame survey consisting of a random list sample of farmers selected from the Farm Register and a random area sample of segments. Enumerators visit the sampled farms in late June and early July. Acreage estimates from the survey are available in early August. Each year about 20% of the segments are changed.

During the growing season, two potato crop area estimates are published. The first, in late June, is based on the mail survey results. The second estimate, in early September, is based on a review of the estimates from all three surveys and discussion with provincial authorities. The date of the second estimate was the target date for generation of a satellite-derived estimate.

## 5. PROJECT REGION

The area for which an estimate was required is located in the upper St. John Valley in New Brunswick. It starts in the south at Woodstock in Carleton County and follows the St. John River for 200 kilometers northwest through Victoria County to Claire in Madawaska County.

The region is heavily wooded, of varied, rolling terrain. There are some problems related to stoniness and drainage. Within the area are 70,000 hectares of improved cropland, of which about 20,000 are usually potatoes. Other major crops are grains, hay and processing vegetables such as peas, broccoli and brussels sprouts. Parcel sizes range from 0.1 hectare seed plots to 40 hectare fields.

## 6. 1980 GROUND SAMPLE AND DATA COLLECTION

The area sample for the A.E.S. in New Brunswick was considered a suitable vehicle for obtaining ground data for interpreting remote sensing data. This sample is selected in two stages. At the first stage, census Enumeration Areas (EA's), which had farm headquarters in them in the 1976 census (called Census Agricultural EA's), were stratified based on their potato acreages, cattle, and pig numbers (1976 census data). Within each stratum, two replicated simple random samples of EA's were selected. Each sampled EA was segmented into identifiable area units of about 5 to 8 square kilometers using maps, and a simple random sample of one in 10 segments was selected per EA. A.E.S. enumerators working in the study region were supplied with old enlarged

aerial photos (scale 1" = 832') for each sampled segment. The photos were obtained from provincial sources. Most of them had been taken in 1976. While contacting the farmers operating land inside the segment, they were required to show the photograph to the farmer and identify on it all potato and corn<sup>3</sup> fields and note their areas as reported by the farmer. Written instructions on procedures to be followed were included in the interviewers' manual and interviewers and supervisors were trained on procedures to be followed.

## 7. ANALYSIS OF REMOTELY SENSED DATA

### 7.1 Previous Work

Work in the same region using 1975 data has been reported elsewhere (Mosher et al. 1978), and a detailed description of the approach is available (Ryerson et al. 1980). In the 1975 work, a test area which contained about twenty percent of the province's potato crop was selected for detailed analysis from the potato growing region. This was supported by ground data collection for the entire test area.

The 125 square kilometer test area and two sub-areas were located on the colour video display screen of the CCRS Image Analysis System (CIAS) (Goodenough, 1979). A very simple supervised training scheme was used to gather the statistics of pixels in three contiguous potato fields in the form of four one-dimensional histograms. A four-dimensional parallelepiped in feature space was generated as defined by the limits of each of the histograms to serve as a decision boundary. All points within the parallelepiped were classified as potatoes, and those outside the region as "other".

One of the major problems was the proper classification of boundary pixels. These present special problems, as they fall on the border of two different

---

<sup>3</sup> Corn fields were also required to be identified since earlier remote sensing work indicated that corn was a confusion crop for potatoes (Ryerson et al. 1980).



fields. Their reflectance is a function of the amount of each field within the pixel and the reflectance of each cover material in the two fields. To attempt to compute the percentage of each cover type present in such pixels is generally very complicated. However, it was found that by modifying the original decision boundary through adding a second parallelepiped formed by training on a number of boundary pixels which appeared to be in potato fields, reasonable area estimates could be achieved. Selection of the appropriate boundary pixels to be classed as potatoes was on the basis of subjective visual interpretation of the display (data from three of the four spectral bands were merged to form a colour display, with colours simulating those of a colour infrared film).

Less than four hours of CIAS time were required to perform the area estimate for the entire potato belt. Location, display and analysis of the primary and sub-test areas required just over one hour. Selection, display and analysis of the subsequent five subscenes required two and one-half hours, while location of the New Brunswick border and elimination of data from outside the province required another hour.

Compared to total area of potato fields interpreted and measured from low altitude aerial photographs taken at the same time, the 1975 satellite estimates were 95% accurate (i.e., 95% of the estimated true value), in the sub-area containing the training site, 80% accurate in the second sub-area, and 88% accurate over the whole primary test area. On repeated tests using different training fields, the accuracy over the primary test area ranged from 85% to 97%. The province wide accuracy was 84.5%. Some of the error in the provincial estimate resulted from the fact that some potatoes are grown outside the potato belt. Other factors contributing to the error are discussed below.

## 7.2 PROCEDURES TO IMPROVE ESTIMATES

Although the previous work was successful, potential sources of error were identified for applications requiring accuracies greater than say 85%. The



major problems arise in the subjective selection of the potato field boundary pixels, in the handling of small fields, and in the resolution of difficulties with the crops confused with potatoes.

With regard to confusion crops, ideally one should know the spectral reflectances of potatoes and all of its confusion crops throughout the growing season. With such information, it is possible to specify the phenological window during which potatoes can be reliably separated from other crops. Unfortunately, such a data set does not exist, although knowledge of the region's crops and the cultural practices does provide some general guidelines. In this case, based on the field experience of the authors, it was hypothesized that the optimum date for separation of potatoes from other crops in this region would be from mid-July to mid-August. To test this hypothesis and provide an indication of the degree of separability of potatoes from other crops, an analysis was performed of a Landsat MSS Computer Compatible Tape (CCT) acquired over the St. John Valley on August 8, 1975. Figure 1 shows the Landsat band radiance values for potatoes, corn, peas, hay, broccoli, pasture, buckwheat, bare soil and grains. It can be seen from this that potatoes are easily separable from the other crops except for the peas, which are usually harvested by mid to late August. It would therefore appear that the analysis of data collected late in the growing season is likely to lead to the separation and identification of potatoes.

The problem of small fields and boundary pixels can be handled by an approach which uses available ground information over a limited area to produce a more accurate crop area estimate over an extended area. Given the size of the potato growing region here, the area of potatoes within each of ten to fifteen segments is required. Each segment is of the order of five to eight square kilometers in size. The whole area is still classified as well as possible, but the subjective boundary pixel class is not produced. The classification result for each segment is then used, along with the available ground information (the aerial photograph data in 1980, A.E.S. data for future years) to obtain a regression relationship which is applied to the entire area estimate to

produce a revised estimate (Hanuschak et al. 1979). A ratio estimate, based on the total area estimates obtained from satellite and aerial photograph data, is also produced.

### 7.3 GENERATION OF THE 1980 ESTIMATE USING LANDSAT

The generation of a satellite-based potato area estimate can be described as a three part process: ordering of old and new data, pre-location of A.E.S. segment boundaries, and image analysis.

The Landsat data ordered where Digital Image Correction (DICS) Computer Compatible Tapes (CCT's) using the  $\sin x / x$  interpolation for geometric correction (Guertin et al. 1979) and Cal 3 radiometric correction (Ahern and Murphy, 1978). Each CCT covers four 1:50,000 National Topographic System (NTS) map sheets with a resampled square pixel of 50M. Four CCT's are required to cover the region. Existing data were ordered for delivery in May of 1980, while new data were ordered for the appropriate satellite passes from mid-July to mid-August. The ordering process proceeded smoothly for existing data, but was complicated for the real-time data by the failure of Landsat-III. A Landsat-II pass on August 17 was used to create DICS CCT's which were delivered to the analysis facility on August 22, well ahead of schedule.

The pre-location of A.E.S. segments was done in the spring of 1980 using the polygon cursor option on the CCRS Image Analysis System (Goodenough, 1977). A.E.S. segment boundaries were provided by Statistics Canada on 1:50,000 map sheets and on photocopies of the airphoto enlargements given to the A.E.S. enumerators. Although some segments had boundaries which were easy to locate (streams, forested edges, lakes, etc.), others based on political or census boundaries proved very complex. Segments whose boundaries were a combination of major roads and/or major rivers could be located, bounded and stored after less than five minutes work on the enlarged colour display of 128 x 128 50M pixels on a 512 x 512 monitor. The most complex took up to an hour - with an

average of 20-25 minutes. Once located, the segment was stored by specific DICS pixel coordinates so that it could be overlaid on new data as it arrived to locate both training data and inputs for the estimator. Because of the location of some segments on the boundary between two DICS CCT's and other similar problems, a number of segment boundaries were not located in the preparatory phase. Software is now being written to shorten the time required for the entire project, especially the pre-location phase. Use of original photographs in place of photocopies planned for the 1981 project's new segments (15 rotated in for 1981), should also shorten the time required.

Once the 1980 ground data were delivered to the analysis center, potential training sites (based on field size) were selected. Several fields were selected from one segment in the north of the region (near St. André) while several others were selected from a segment in the south (near Hartland).

Upon receipt of the 1980 satellite data, it became a relatively simple task to recall the segment boundaries, overlay them, locate training fields and begin classification using methods described in 7.1. In addition to the selected training areas, another group of large fields was selected as were areas of known potato fields which appeared brighter red on the monitor than those in the training set. As classification results were available, crop areas were recorded for each 512 x 512 pixel subscene and for each of seventeen A.E.S. segments.

There were four problems encountered during the classification; one involving imputation of crop under scattered cloud and cloud shadow and the other three related to confusions. The method of imputation of potatoes under cloud was quite straightforward. It was assumed that the percent area of crop under the clouds was similar to the area of crop in an adjacent "like-appearing" region. A simple formula was used to determine potato area under cloud, PC:

$$PC = \frac{P_M}{T_A - C_A} \times C_A$$

where:  $P_M$  = potato area measured in the cloud free region;  $C_A$  = area in cloud and shadow in the region and  $T_A$  = total area in the region. These areas were incorporated into the total satellite estimate, no A.E.S. segments under cloud were included in the ratio or regression analyses. The problems with confusions were, for the most part, solved through careful modification of the classification parameters. In one case, an unknown form of widely scattered natural regrowth in forest cut-overs was confused. In another case, hay fields with regularly spaced piles of stone were similar to potatoes. Only a few fields of clover in one segment and some of the fairways of a golf course in another remained as confusions after modifying the classification.

The areas calculated for the region are presented and discussed in more detail below.

## 8. RESULTS AND ANALYSIS OF THE 1980 NEW BRUNSWICK POTATO PROJECT

### 8.1 INTRODUCTION

The analysis of data from the 1980 New Brunswick potato project was done at the regional, segment, and field levels. Two types of estimates were obtained for the total potato area of the test region in the St. John Valley using satellite data and ground verified measures from high altitude aerial photograph data (the aerial photography data were obtained and analysed by CCRS). The estimates and their variances were then compared to other estimates from Statistics Canada surveys in New Brunswick. Segment potato acreages reported by the A.E.S. and by satellite were then compared to the aerial photograph acreages (which are considered here to be closest to the actual values) to determine the strength of their relationship at the segment level. This analysis was complemented by examining the variation in A.E.S. reporting of field acreages for each interviewer's assigned area.



## 8.2 ESTIMATION OF POTATO AREAS USING SATELLITE DATA

A ratio and a regression estimate of the total potato area in the test region in New Brunswick were obtained using satellite data and the aerial photograph segment data. Estimates, along with their variances, were calculated using the A.E.S. sample design. The A.E.S. in New Brunswick is a multiple-frame stratified replicated two-stage sample of segments, designed to give accurate estimates of various items at the provincial level (see sections 4 and 6). Since the A.E.S. strata did not coincide with the test region boundaries, the technique of post-stratification was used for estimation, treating the EA's as a with-replacement-sample. Segments with missing satellite data were not included in the sample, nor was one outlier (see Figure 3). These estimates were based on 40 segments. Finally, since A.E.S. enumerators did not always collect data on all farms inside the segment (see 8.3), aerial photograph segment acreages were used for estimation.

Let the label  $x$  represent the reported satellite potato data and  $y$  represent aerial photograph data. The ratio and regression estimates of  $Y$ , the total potato area in the test region were calculated as:

$$\hat{Y}_{\text{Ratio}} = \hat{R} X = \frac{\hat{Y}}{\hat{X}} X, \text{ and}$$

$$\hat{Y}_{\text{Reg.}} = \hat{Y} + \hat{B} (X - \hat{X}),$$

where  $\hat{Y} = \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$  is the design estimate of  $Y$ ;

$\hat{X}$  is the design estimate of  $X$ , the total uncorrected satellite potato area in the test region.  $\hat{X}$  is obtained by substituting  $x_{hij}$  for  $y_{hij}$  in the formula for  $\hat{Y}$ ;



$y_{hij}$  and  $x_{hij}$  are the observed values for the  $j$ th selected segment in the  $i$ th selected first-stage unit (EA) from stratum  $h$ ;

$N_h$  and  $n_h$  are the total and sampled number of EA's in stratum  $h$ ,  $h=1, \dots, L$ ;

$M_{hi}$  and  $m_{hi}$  are the total and sampled number of second-stage units (segments) in the  $i$ th selected EA of stratum  $h$ ;

$\hat{R}$  is an estimate of  $Y/X$ ; and,

$\hat{B} = \text{cov}(\hat{Y}, \hat{X}) / \text{var}(\hat{X})$  is the linear regression coefficient.

The variance estimates of  $\hat{Y}_{\text{Ratio}}$  and  $\hat{Y}_{\text{Reg.}}$  are given by:

$$v(\hat{Y}_{\text{Ratio}}) = v(\hat{Y}) - 2\hat{R} \text{cov}(\hat{Y}, \hat{X}) + \hat{R}^2 v(\hat{X}), \text{ and}$$

$$v(\hat{Y}_{\text{Reg.}}) = v(\hat{Y}) - 2\hat{B} \text{cov}(\hat{Y}, \hat{X}) + \hat{B}^2 v(\hat{X}),$$

$$\text{where } v(\hat{Y}) = \sum_{h=1}^L \frac{N_h^2}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left[ \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} - \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \right]^2$$

and  $v(\hat{X})$  is calculated by substituting  $x_{hij}$  for  $y_{hij}$  in  $v(\hat{Y})$ .

It can be seen that  $\hat{B} = \text{cov}(\hat{Y}, \hat{X}) / v(\hat{X})$  is the value of  $B$  which minimizes  $v(\hat{Y}_{\text{Reg.}})$ .

Table 1 shows the ratio and regression estimates, along with other Statistics Canada estimates of the potato area in New Brunswick. The ratio and regression estimates, pro-rated to the provincial level, are both very close to the Statistics Canada published figure of 52,000 acres. There is very little difference between the two estimates, and they both have the same coefficient of variation. In order to give an idea of the gain in efficiency obtained by the ratio estimate, the variance of the ratio estimate was as low as one-fifth of that of  $\hat{Y}$ , design estimate of  $Y$  based on the post-stratified design (area sample only). It may be noted that the C.V.'s of the ratio and regression esti-

mates are of the same order as that of the A.E.S. multiple-frame estimate, although the latter is based on a larger sample size.

### 8.3 COMPARISON OF DATA AT THE SEGMENT AND FIELD LEVEL

Figures 2 and 3 show plots of segment potato acreages, as reported by the A.E.S. enumerators and by satellite, against aerial photograph acreages. Not all sample segments were used in the analysis. Satellite data were missing for 16 segments due to cloud cover and image location. Eight A.E.S. segments, containing survey non-respondents and very large farms whose field data were not to be collected by the enumerators, were not used. The two outliers in the plots are not used in the calculations here and in Section 8.2.

The plots show a strong linear relationship between A.E.S. and aerial photograph data, as well as between satellite and aerial photograph data, at the segment level, with correlations<sup>4</sup> of .991 and .968 respectively<sup>5</sup>. There is a tendency for both the A.E.S. enumerators and the satellite classification to underestimate the acreages. This is less marked for the A.E.S. Causes of discrepancies of satellite acreages were explained in Section 7. Some segments with little or no potatoes were over-estimated because of confusion crops. (The satellite outlier had confused a large hay field with rocks for a potato field - this error could have been removed by modifying the classification). One major cause of A.E.S. underestimation was that the interviewers did not enumerate some farms in the segment because of the multiple-frame procedures (these include farms which appeared on the list frame as well as farms that had land in more than one sample segment). Specific instructions have been written up in the 1981 A.E.S. field procedures to ensure that all farms in the segment are enumerated for their potato acreages next year when the test is to be repeated. This is expected to bring the A.E.S. reported acre-

---

<sup>4</sup> The correlations were estimated using the sample design weights.

<sup>5</sup> A plot of satellite data against the A.E.S. data looked very similar to Figure 3, with a correlation of .957.

ages closer to the aerial photograph acreages. The strong relationship between the A.E.S. and the aerial photograph data for the sampled segments is encouraging and supports making adjustments to the satellite estimates using the data collected by the A.E.S. enumerators.

Another cause of A.E.S. discrepancy with aerial photograph segment data was the mis-reporting of field boundaries and field sizes. A.E.S. field acreages were obtained by interviewing farm operators, and thus, were frequently reported in multiple of five acres. Plots of A.E.S. reported field acreages against aerial photograph acreages by interviewer assignment areas indicate that there may be a difference in field reporting between the assignment areas. This could be caused by the interviewers themselves, but also by other factors such as geographic location and structures of fields within sample segments, or by variation in reporting errors. More variation was observed in the region east of the St. John River, where average field sizes were generally larger. The relationship between the A.E.S. data and aerial photograph data was stronger at the segment level than at the field level.

## 9. SUMMARY AND CONCLUSION

Satellite data were used in 1980 to generate a highly accurate potato area estimate for the three major potato producing counties of New Brunswick. Through the project described here, refinements have been identified for field procedures and analysis methods which should provide even more accurate estimates of crop area with satellite data, reduce respondent burden, and provide detailed spatial information previously available only in Census years.

## 10. ACKNOWLEDGEMENTS

The authors would like to acknowledge the technical advice and editorial comments of Mrs. N. Chinnappa of Statistics Canada and of Drs. J. Cihlar and F. Ahern of the Canada Centre for Remote Sensing.

## 11. REFERENCES

- [1] Ahern, F.J. and J. Murphy, 1978, "Radiometric Calibration and Correction of Landsat 1, 2, and 3 MSS Data", CCRS Research Report 78-4, Energy, Mines and Resources, Ottawa, Canada.
- [2] Brown, R.J., Ahern, F.J., Ryerson, R.A., Thomson, K.P.B., Goodenough, D.G., McCormick, J.A. and Teillet, P.M., 1980, "Rapeseed: Guidelines for Operational Monitoring", 6th Canadian Symposium on Remote Sensing, Halifax, Nova Scotia, Canada, 321-330.
- [3] Cochran, W.G., 1977, Sampling Techniques (3rd Edition), John Wiley and Sons Inc., New York.
- [4] Goodenough, D.G., 1979, "The Image Analysis System (CIAS) at the Canada Centre for Remote Sensing". Canadian Journal of Remote Sensing 5(1) May 1979, 3-17.
- [5] Guertin, F.G., T.J. Butlin and R.G. Jones, 1979, "Correction geometrique des images Landsat au Centre canadien de télédétection", Canadian Journal of Remote Sensing, 5(2), December 1979.
- [6] Hanuschak, G., R. Sigman, M. Craig, M. Ozga, R. Luebbe, P. Cook, D. Kleweno, and C. Miller, 1979, "Crop - Area Estimates with Landsat: Transition from Research and Development to Timely Results", Proceedings of the Fifth Machine Processing of Remotely Sensed Data Symposium, Lafayette, Indiana.
- [7] Mosher, P.N., R.A. Ryerson and W.M. Strome, 1978, "New Brunswick Potato Area Estimation from Landsat", 12th International Symposium on Remote Sensing of Environment, Manila, Philippines, April 1978, 1415-1419.
- [8] Ryerson, R.A., P. Mosher, V. Wallen and N. Stewart, "Three Tests of Agricultural Remote Sensing in Eastern Canada: Results, Problems and Prospects", Canadian Journal of Remote Sensing, 5(1) 1979, 53-66.
- [9] Ryerson, R.A., P.N. Mosher and J. Harvie, 1980, "Potato Area Estimation Using Remote Sensing Methods", CCRS Users Manual 80-2, Energy, Mines and Resources, Ottawa, February 1980.

TABLE 1

SURVEY ESTIMATES OF THE TOTAL POTATO AREA FOR THE TEST REGION  
AND FOR THE PROVINCE OF NEW BRUNSWICK<sup>1</sup>

Survey and/or Estimate	Test Region Estimate (Acres)	New Brunswick Estimate (Acres)	Coefficient of Variation (%)
Satellite			
Uncorrected	42,354	N/A	N/A
Ratio	49,504	51,524	5.5
Regression	49,115	51,119	5.5
Mail Survey	N/A	50,800	N/A
A.E.S. Multiple Frame	N/A	53,854	5.2
O.P.Y.S.	47,203	49,129	4.59
Statistics Canada Published (Sept. 5)	N/A	52,000	N/A

<sup>1</sup> The test region, composed of the counties of Carleton, Madawaska, and Victoria, accounts for about 96.08% of the total potato area of New Brunswick.



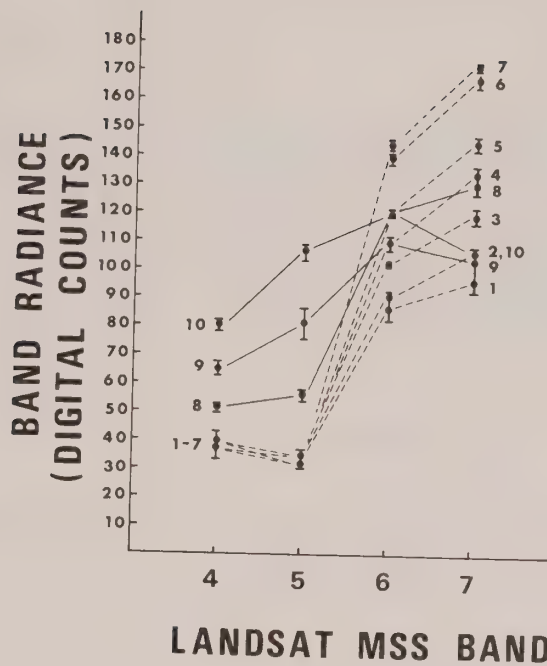
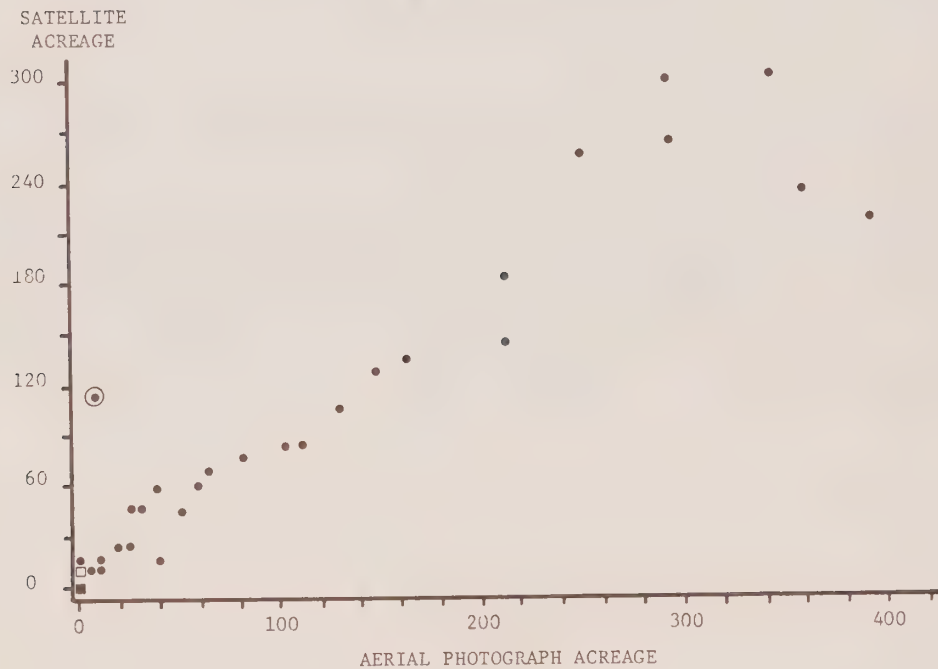
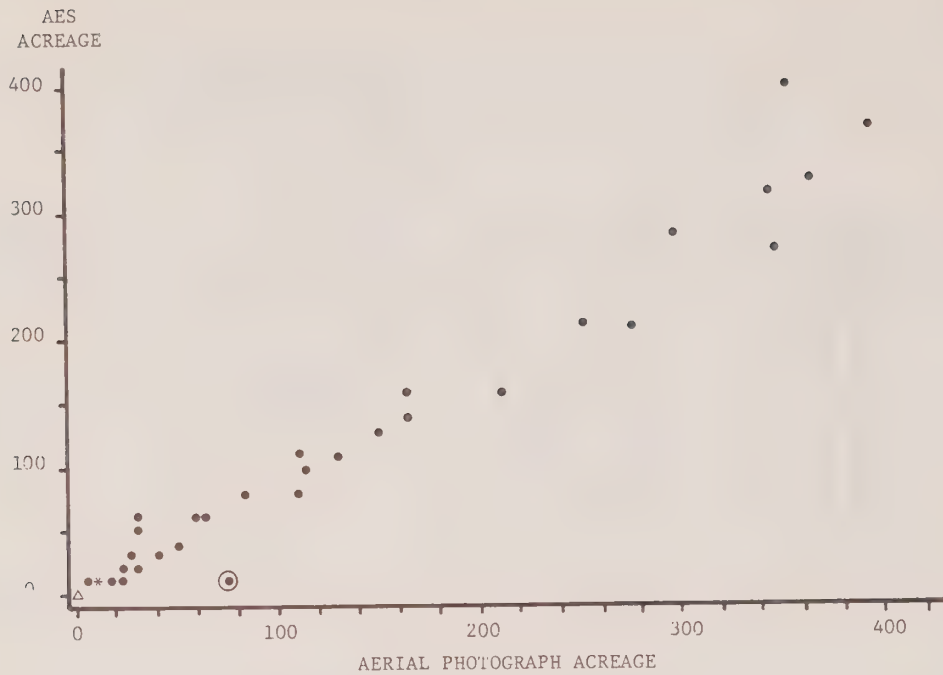


FIGURE 1 Comparison of Satellite Band Radiances for Various Crops.

Landsat MSS band radiance (August 8, 1975) for grains (1), buckwheat (2), pasture (3), hay (4), corn (5), peas (6), potatoes (7), broccoli (8), bare soil (9), weeds (10). The bands are numbered 4 to 7. The relative distances between the digital counts in each column indicate how separable the crops are for the given band.



FIGURES 2 and 3 - Plots of reported A.E.S. and satellite potato acreages vs. low altitude aerial photograph acreages for sampled segments in New Brunswick. The circled observations represent outliers.

Legend: . = 1 obs., □ = 2 obs., \* = 3 obs., ■ = 11 obs., and Δ = 15 obs.

## SAMPLING ON TWO OCCASIONS WITH PPSWOR

G.H. Choudhry and Jack E. Graham<sup>1</sup>

A theory of sampling on two occasions with unequal probabilities and without replacement is presented. Fellegi's (1963) method, which yields the same selection probabilities for a given unit on each occasion, is used to select the units for the rotation sample. The variances of composite estimators of the population total on the second occasion are developed. Numerical results are presented for small sample sizes and efficiency comparisons are made with a competing strategy.

## 1. INTRODUCTION

In surveys of a repetitive nature there are advantages to using a partial replacement sampling scheme both from the point of view of efficiency of estimation as well as reduction of respondent's burden. Essentially, after each sampling occasion a fraction of the units is rotated out of the sample and is replaced by a fresh subsample from the population. The literature abounds with discussions of sampling procedures and estimators when sampling on two or more occasions with equal probability. But of particular practical importance is the situation where units are selected on a given occasion with unequal probabilities. Thus, consider a finite population of  $N$  units  $\{1, 2, \dots, N\}$  and two sampling occasions 1 (the previous occasion) and 2 (the current occasion). Let  $y_{1i}$  and  $y_{2i}$  denote the values of a characteristic  $y$  borne by the  $i$ -th unit on occasions 1 and 2 and let  $Y_1$  and  $Y_2$  denote the respective population totals. A size measure  $x_i$  is known for each of the units in the population.

Raj (1965) considered the following pps (probabilities proportional to size) sampling scheme: on the first occasion a sample  $s$  of size  $n$  is selected with probabilities  $p_i$  proportional to the  $x_i$  values and with replacement

<sup>1</sup> G.H. Choudhry, Census & Household Survey Methods Division, Statistics Canada and Jack E. Graham, Carleton University.

(wr). On the second occasion a simple random sample  $s_1$  of  $m$  units is selected from  $s$  without replacement (wor) and an independent pps sample  $s_2$  of  $u = n-m$  is selected wr from the entire population.  $Y_1$  and  $Y_2$  are then respectively estimated by

$$\hat{Y}_1 = \sum_s y_{1i} / (np_i) \text{ and } \hat{Y}_{2R}^* = Q^* \hat{Y}_{2u} + (1-Q^*) \hat{Y}_2',$$

$$\text{where } \hat{Y}_{2u} = \sum_{s_2} y_{2i} / (up_i), \hat{Y}_2' = \hat{Y}_1 + \sum_{s_1} (y_{2i} - y_{1i}) / (mp_i),$$

and  $Q^*$  is a weight,  $0 \leq Q^* \leq 1$ .

The minimum variance of  $\hat{Y}_{2R}^*$  was developed under the assumption that

$$v_{pps}(y_t) = \sum_{i=1}^N p_i (y_{ti}/p_i - Y_t)^2$$

is the same for  $t = 1$  and  $2$ .

The problem of sampling with ppswor on one occasion has attracted considerable attention in the literature. A major difficulty lies in the specification of feasible procedures which lead to specified probabilities at each and every draw. Fellegi (1963) has proposed a method such that the probability that unit  $i$  is selected on each of the  $n$  draws is  $p_i$  by determining  $n-1$  sets of "working probabilities". This is an extremely desirable feature for rotating samples where it is essential that the usual pps estimator be unbiased for  $Y_2$ ; this will not be true for any partial replacement design that does not feature a constant  $p_i$  for each of the  $n$  draws. The calculations inherent in Fellegi's scheme have, until recently, been prohibitive for  $n > 2$ . Choudhry (1981) has developed an iterative procedure for implementing Fellegi's scheme and prepared a computer program to evaluate the working probabilities when  $n \leq 5$ . Although the convergence is fast in terms of the number of iterations, the amount of computation increases at the rate  $N^n$ .

The program also computes the joint probabilities for the inclusion of both units  $i$  and  $j$  in the sample for variance calculation purposes.

Rao, Hartley and Cochran (1962) devised the "random group method" for selecting a sample with ppswor. The population of  $N$  units is split into  $n$  groups of sizes  $N_1, N_2, \dots, N_n$  where  $\sum N_h = N$  and a sample of one unit is drawn independently from each group with probabilities proportional to the  $p_i$ 's. Ghangurde and Rao (1969) extended the random group method to sampling on two occasions. For simplicity, the  $N$  units were split into  $n$  groups each of size  $N/n$  (assumed to be an integer). On occasion 1, one unit is drawn from each random group as above, giving a sample  $s$  of  $n$  units. On the second occasion a simple random sample  $s_1$  of  $m$  matched units is selected from the  $n$  units wor and an independent sample  $s_2$  of  $u=n-m$  units is drawn from the whole population by the method used in obtaining  $s$ . They form a composite estimator  $\hat{Y}'_{2G}$  of  $Y_2$  and obtain its minimum variance under an optimum choice of the weight  $Q$ . The optimum value of  $\lambda = m/n$  is then determined. The authors remarked that it would likely be more efficient to select  $s_2$  from the  $N-n$  units in the population that are not included in  $s$ .

Chotai (1974) modified the Ghangurde-Rao (G-R) design on the second occasion; the  $n$  units in  $s$  are split at random into  $m$  groups of size  $n/m$  (assumed to be an integer). One unit is selected from each of the  $m$  groups with probability proportional to  $p_i$ , yielding a sample  $s_1$ . A sample  $s_2$  is obtained as in the G-R method. The optimum variance of his composite estimator  $\hat{Y}'_{2c}$  is derived, the optimum  $\lambda$  determined and relative efficiency comparisons of  $\hat{Y}'_{2c}$  with respect to G-R's and Raj's optimal estimators are made.  $\hat{Y}'_{2c}$  was found to be always more efficient than  $\hat{Y}^*_{2R}$  and, in many cases,  $\hat{Y}'_{2G}$  as well. A brief discussion of the case when  $n/m$  is not an integer is provided. It is worth noting that because  $\lambda$  is not a continuous function, the optimum  $\lambda$  should really be determined using integer programming methods. In what follows a sampling strategy is developed which often results in greater gains in efficiency over ppswor sampling than previously proposed schemes.



## 2. SAMPLING STRATEGY

### 2.1 SAMPLING PROCEDURE

From the population of  $N$  units,  $(1, 2, \dots, N)$ , select a sample of  $n+u$  units,  $u < n$ , draw by draw and without replacement using Fellegi's Method such that the probability of selecting the  $i$ -th unit at each draw is  $p_i$ ,  $i=1, 2, \dots, N$ ,  $\sum p_i = 1$ . On the first of the two occasions, the first  $n$  units are observed from the  $n+u$  selected; on the second occasion the first  $u$  units are dropped from the sample and the unused set of  $u$  units is rotated into the sample. Thus  $m = n-u$  units are observed on both occasions. The  $n$  units observed on the first occasion are referred to as  $s$ , those units observed on both occasions as  $s_1$  (where  $s_1 \subset s$ ) and the set of unmatched units observed only on the second occasion as  $s_2$ . Note that Fellegi's scheme guarantees that the selection probabilities for a given unit  $i$  are the same on each draw and hence the same on both occasions. By restricting his attention to a sub-class of non-homogeneous linear model-design unbiased estimators, Chaudhuri (1980) has shown that the foregoing sampling scheme yields an optimal strategy. This is a further motivation for using Fellegi's method.

### 2.2 ESTIMATION THEORY

In what follows, composite estimators of  $Y_2$ , the current occasion total, are proposed and their variances determined using an indicator variable approach.

Let  ${}_r a_i = 1$  if unit  $i$ ,  $i=1, 2, \dots, N$ , is selected at draw  $r$ ,  $r=1, 2, \dots, n+u$  and  ${}_r a_i = 0$  otherwise. Since the expectation of  ${}_r a_i$  is  $p_i$ , an unbiased estimator of the first occasion population total  $Y_1$  is

$$\hat{Y}_1 = \frac{1}{n} \sum_{r=1}^n \sum_{i=1}^N {}_r a_i y_{1i} / p_i.$$

Then 
$$\hat{Y}'_2 = \hat{Y}_1 + \frac{1}{m} \sum_{r=u+1}^n \sum_{i=1}^N a_{ri} (y_{2i} - y_{1i}) / p_i$$

is an unbiased estimator of the second occasion total  $Y_2$ . An unbiased estimator of  $Y_2$  based on the current observations is

$$\hat{Y}_2 = \frac{1}{n} \sum_{r=u+1}^{u+n} \sum_{i=1}^N a_{ri} y_{2i} / p_i$$

A composite estimator of  $Y_2$  is the weighted sum

$$\hat{Y}_{2c} = Q\hat{Y}'_2 + (1-Q)\hat{Y}_2,$$

where  $0 \leq Q \leq 1$ .

The variance of  $\hat{Y}_{2c}$ ,  $\text{Var}(\hat{Y}_{2c}) = Q^2 \text{Var}(\hat{Y}'_2) + (1-Q)^2 \text{Var}(\hat{Y}_2) + 2Q(1-Q)\text{Cov}(\hat{Y}'_2, \hat{Y}_2)$ , is derived by using the following properties of the indicator variable  $a_{ri}$ :

$$\text{Var}(a_{ri}) = p_i(1-p_i), (i=1,2,\dots,N, r=1,2,\dots,n+u),$$

$$\text{Cov}(a_{ri}, a_{ti}) = -p_i^2, (r \neq t),$$

$$\text{Cov}(a_{ri}, a_{rj}) = -p_i p_j, (i \neq j),$$

$$\text{Cov}(a_{ri}, a_{tj}) = E(a_{ri} \cdot a_{tj}) - p_i p_j, \text{ otherwise,}$$

where  $E(\cdot)$  denotes the expected value with respect to the probability design.

Now  $E(a_{ri} \cdot a_{tj}) = P(a_{ri} \cdot a_{tj} = 1) = P(a_{ri} = 1, a_{tj} = 1)$

where  $P(\cdot)$  denotes probability.

Let  $\Sigma_{(k-2; i, j)}$  denote summation over all possible ordered  $(k-2)$ -tuples of different units  $\{i_1, i_2, \dots, i_{r-1}, i_{r+1}, \dots, i_{k-2}, i_{k-1}\}$  included in the sample from the first  $k$  draws selected from the  $N-2$  units in the set  $\{1, 2, \dots, i-1, i+1, \dots, j-1, j+1, \dots, N\}$  such that the  $i$ -th unit is selected at draw  $r$  and the  $j$ -th unit at draw  $k$ . There are  $(N-2)(N-3)\dots(N-k+1)$  terms involved in the summation.

As in Fellegi (1963), let  $\{p_i(\ell); i=1,2,\dots,N\}$  be the set of "working probabilities" for selecting a unit at draw  $\ell$ ,  $\ell=1,2,\dots,n+u$ . For draws  $k$  and  $r$  with  $k > r$ ,

$$E(a_{ri} \cdot a_{kj}) = \sum_{(k-2;i,j)} p_{i_1}(1) \frac{p_{i_2}(2)}{1-p_{i_1}(2)} \dots \frac{p_{i_{r-1}}(r-1)}{1-\sum_{\ell=1}^{r-2} p_{i_\ell}(r-1)} \times$$

$$\times \frac{p_{i_r}(r)}{1-\sum_{\ell=1}^{r-1} p_{i_\ell}(r)} \times \frac{p_{i_{r+1}}(r+1)}{1-\sum_{\ell=1}^{r-1} p_{i_\ell}(r+1)-p_{i_r}(r+1)} \times$$

$$\dots \times \frac{p_{i_k}(k)}{1-\sum_{\ell=1}^{r-1} p_{i_\ell}(k) - p_{i_r}(k) - \sum_{\ell=r+1}^{k-1} p_{i_\ell}(k)} .$$

Now

$$\text{Var}(\hat{Y}'_2) = \frac{1}{n^2} \text{Var} \left[ \sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i \right] + \frac{1}{m^2} \text{Var} \left[ \sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right]$$

$$+ \frac{2}{mn} \text{Cov} \left[ \sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i, \sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right]$$

Using the previously cited properties of the indicator variables  $r a_i$  it may be verified that

$$\frac{1}{n^2} \text{Var} \left[ \sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i \right] = \frac{1}{n} \sum_{i=1}^N p_i z_{1i}^2 + \frac{1}{n^2} \sum_{i \neq j} P(i,j \in S) z_{1i} z_{1j} - Y_1^2 ,$$

$$\frac{1}{m^2} \text{Var} \left[ \sum_{r=u+1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right] = \frac{1}{m} \sum_{i=1}^N p_i (z_{2i} - z_{1i})^2$$

$$+ \frac{1}{m^2} \sum_{i \neq j} \sum P(i, j \in s_1) (z_{2i} - z_{1i})(z_{2j} - z_{1j}) - (Y_2 - Y_1)^2,$$

where  $z_{tj} = y_{tj}/p_j$ ,  $t=1,2$  and  $n-u=m$ .

Also,

$$\begin{aligned} & \frac{1}{mn} \text{Cov} \left[ \sum_{r=1}^n \sum_{i=1}^N r a_i y_{1i} / p_i, \sum_{u=1}^n \sum_{i=1}^N r a_i (y_{2i} - y_{1i}) / p_i \right] \\ &= \frac{1}{n} \sum_{i=1}^N p_i z_{1i} (z_{2i} - z_{1i}) + \frac{1}{mn} \sum_{i \neq j} \sum P(i \in s, j \in s_1) z_{1i} (z_{2j} - z_{1j}) - Y_1 (Y_2 - Y_1). \end{aligned}$$

Combining the foregoing 3 terms gives

$$\begin{aligned} \text{Var}(\hat{Y}'_2) &= \sum_{i=1}^N p_i \left[ \frac{z_{2i}^2}{n} + (z_{2i} - z_{1i})^2 \left( \frac{1}{m} - \frac{1}{n} \right) \right] + \sum_{i \neq j} \left[ \frac{P(i, j \in s)}{n^2} z_{1i} z_{1j} \right. \\ &+ \frac{P(i, j \in s_1)}{m^2} (z_{2i} - z_{1i})(z_{2j} - z_{1j}) \\ &+ \left. \frac{2P(i \in s, j \in s_1)}{nm} z_{1i} (z_{2j} - z_{1j}) \right] - Y_2^2. \quad (1) \end{aligned}$$

Also,

$$\text{Var}(\hat{Y}_2) = \frac{1}{n} \sum_i p_i z_{2i}^2 + \frac{1}{n^2} \sum_{i \neq j} \sum P(i, j \in s^*) z_{2i} z_{2j} - Y_2^2, \quad (2)$$

where  $s^*$  is the set of  $n$  units observed on the second occasion, and

$$\text{Cov}(\hat{Y}_2, \hat{Y}'_2) = \sum_{i \neq j} \left[ \frac{P(i \in s, j \in s^*)}{n^2} z_{1i} z_{2j} + \frac{P(i \in s_1, j \in s^*)}{nm} (z_{2i} - z_{1i}) z_{2j} \right]$$

$$+ \frac{1}{n} \sum_i p_i z_{2i} (z_{2i} - \frac{u}{n} z_{1i}) - Y_2^2 \quad (3)$$

Expressions (1), (2) and (3), when combined, yield  $\text{Var}(\hat{Y}_{2c})$ .

The optimum value of the weight  $Q$  which minimizes  $\text{Var}(\hat{Y}_{2c})$  is

$$Q_{\text{opt}} = [\text{Var}(\hat{Y}_2) - \text{Cov}(\hat{Y}_2', \hat{Y}_2)] / [(\text{Var}(\hat{Y}_2') + \text{Var}(\hat{Y}_2) - 2 \text{Cov}(\hat{Y}_2', \hat{Y}_2))].$$

The corresponding minimum variance is

$$\text{Var}(\hat{Y}_{2c}) = [\text{Var}(\hat{Y}_2') \cdot \text{Var}(\hat{Y}_2) - (\text{Cov}(\hat{Y}_2', \hat{Y}_2))^2] / [\text{Var}(\hat{Y}_2') + \text{Var}(\hat{Y}_2) - 2 \text{Cov}(\hat{Y}_2', \hat{Y}_2)].$$

An alternative composite estimator  $\hat{Y}_{2c}^*$  of  $Y_2$  is

$$\hat{Y}_{2c}^* = Q^* \hat{Y}_2' + (1 - Q^*) \hat{Y}_{2u},$$

where

$$\hat{Y}_{2u} = \frac{n+u}{\sum_{r=n+1}^N} \sum_{i=1}^N r a_i y_{2i} / (u p_i).$$

The variance of  $\hat{Y}_{2c}^*$  is found by combining (1) with

$$\text{Var}(\hat{Y}_{2u}) = \frac{1}{u} \sum_i p_i z_{2i}^2 + \frac{1}{u^2} \sum_{i \neq j} \sum P(i, j \in s_2) z_{2i} z_{2j} - Y_2^2,$$

$$\text{Cov}(\hat{Y}_2', \hat{Y}_{2u}) = \frac{1}{nu} \sum_{i \neq j} \sum P(i \in s, j \in s_2) z_{1i} z_{2j}$$

$$+ \frac{1}{nu} \sum_{i \neq j} \sum P(i \in s_1, j \in s_2) (z_{2i} - z_{1i}) z_{2j} - Y_2^2.$$



### 2.3 SPECIAL CASE

As a check on the calculations, consider the case of simple random sampling without replacement.

Then  $\hat{Y}'_2 = N(\bar{y}_1 + (\bar{y}_{2m} - \bar{y}_{1m}))$  where  $\bar{y}_1, \bar{y}_{1m}$  are, respectively, the sample means based on all the sampled units and all matched units on the first occasion, and  $\bar{y}_{2m}$  is the sample mean based on all matched units on the second occasion.

A direct evaluation gives

$$\text{Var}(\hat{Y}'_2) = N^2 \left[ \left( \frac{1}{m} - \frac{1}{N} \right) (S_1^2 - 2S_{12}) + \left( \frac{1}{m} - \frac{1}{N} \right) S_2^2 \right]$$

where, e.g.,

$$S_{12} = \frac{N}{\sum_{i=1}^N} (y_{1i} - \bar{y}_1) (y_{2i} - \bar{y}_2) / (N-1).$$

This agrees with the result given by (1) with  $p_i = 1/N$  and  $P(i, j \in s) = n(n-1)/N(N-1)$  ( $i \neq j$ ).

Also, under simple random sampling,  $\hat{Y}_2 = N\bar{y}_2$  (where  $\bar{y}_2$  is the sample mean based on all  $n$  sampled units on the second occasion) with variance

$$\text{Var}(\hat{Y}_2) = N(N-n)S_2^2.$$

An evaluation of  $\text{Var}(\hat{Y}_2)$  from (2) gives the same result. Finally,

$$\text{Cov}(\hat{Y}'_2, \hat{Y}_2) = -NS_2^2$$

from either a direct evaluation or from (3). Similarly,  $\text{Var}(\hat{Y}_{2c}^*)$  may also be

checked.

### 3. NUMERICAL EXAMPLES

The composite estimators  $\hat{Y}_{2c}$  and  $\hat{Y}_{2c}^*$  with their optimum  $Q$  and  $Q^*$  values which minimize their respective variances are compared in efficiency with the pps estimator  $\hat{Y}_2$  which is based on the current occasion information only. Because closed forms for  $\text{Var}(\hat{Y}_{2c})$  and  $\text{Var}(\hat{Y}_{2c}^*)$  are not available to permit analytic comparisons to be made, small populations of variate values were employed to affect these contrasts. (The populations studied were necessarily small, like those one might encounter in stratified sampling, since differential effect of sampling with and without replacement is evident only when the sampling fractions are not negligible). Four rotation sampling plans were applied to each population:  $(n,m) = (2,1), (3,2), (3,1)$  and  $(4,3)$  where  $n$  is the number of units in the sample on each occasion and  $m$  is the number of units in the sample common between the two occasions. Two of the populations are given in Murthy (1967) where his single population of 34 villages was subdivided into two populations of sizes 16 and 17 (one outlier unit being discarded). The size measure characteristic is  $x$  = cultivated acreage in 1961 with  $y_1$  and  $y_2$  being the acreage under wheat in 1963 and 1964 respectively. A third population is a set of 14 farms in the province of Saskatchewan with  $x$  = 1980 farm acreage and  $y_1$  and  $y_2$  the 1980 and 1981 cropland acreages respectively. Two additional real data sets relating to populations of sizes 15 and 16 respectively are also analyzed.

Table 1 reports the relative efficiencies of  $\hat{Y}_{2c}$  and  $\hat{Y}_{2c}^*$  with respect to  $\hat{Y}_2$  for each of these 5 populations and 4 sampling plans. A crucial parameter in each comparison is the correlation  $\rho_z$  between  $z_{1i} = y_{1i}/p_i$  and  $z_{2i} = y_{2i}/p_i$ :

$$\rho_z = \frac{\sum_{i=1}^N p_i z_{1i} z_{2i} - Y_1 Y_2}{\sqrt{\sum_{i=1}^N p_i z_{1i}^2 - Y_1^2} \sqrt{\sum_{i=1}^N p_i z_{2i}^2 - Y_2^2}} .$$

The populations studied yielded  $\rho_z$  values ranging from 0.940 to 0.213. The optimum Q and Q\* values are also cited. An investigation of the efficiencies of  $\hat{Y}_{2c}$  and  $\hat{Y}_{2c}^*$  under non-optimum choice of Q and Q\* is planned.

We note the following from these empirical studies: (1) The optimum Q values tend to be larger when  $\rho_z$  is large and as  $\rho_z$  decreases, the optimum Q tends to decrease in both  $\hat{Y}_{2c}$  and  $\hat{Y}_{2c}^*$ . (2) The optimum Q value for  $\hat{Y}_{2c}^*$  always exceeds that for  $\hat{Y}_{2c}$ . (3) As  $\rho_z$  decreases, the efficiency of  $\hat{Y}_{2c}$  with respect to  $\hat{Y}_2$  decreases (as expected), approaching unity as a lower bound under an optimum choice of Q. On the other hand, no such distinct behaviour for  $\hat{Y}_{2c}^*$  is evident since  $\text{Var}(\hat{Y}_{2c}^*)$  is not a monotone function of  $\rho_z$ . For small  $\rho_z$  values, small efficiency gains and losses relative to  $\hat{Y}_2$  are both recorded. (4)  $\hat{Y}_{2c}^*$  is more efficient than  $\hat{Y}_{2c}$  for larger  $\rho_z$  values whereas  $\hat{Y}_{2c}$  is more efficient than  $\hat{Y}_{2c}^*$  for smaller  $\rho_z$  values. (5) If  $\lambda = m/n$  is small, e.g., the (3,1) plan, then large efficiency gains using  $\hat{Y}_{2c}^*$  in preference to  $\hat{Y}_2$  result for large  $\rho_z$ 's. For smaller  $\rho_z$  values, the (4,3) plan yields the largest gains using  $\hat{Y}_{2c}$ ; the other three schemes give about the same gains.

It is worth remarking that even if one has a good correlation between the  $y_{1i}$  and  $y_{2i}$  values, composite estimation using  $\hat{Y}_{2c}^*$  can still lead to efficiency losses compared to the use of the pps estimator  $\hat{Y}_2$  based on current occasion data only. The critical factor is the correlation  $\rho_z$  between the  $z_{1i}$  and the  $z_{2i}$  values. One cannot lose using  $\hat{Y}_{2c}$  under an optimum choice of  $Q$  for  $\hat{Y}_{2c} = \hat{Y}_2$  with  $Q = 0$ .

Table 2 provides the relative efficiencies of  $\hat{Y}_{2c}$  and  $\hat{Y}_{2c}^*$  in the ppswr design with the estimator  $\hat{Y}_{2R}^*$  used by Raj (1965) in his ppswr design described earlier. For more valid comparisons, it was not assumed that  $V_{pps}(y_t)$  was the same for occasions  $t = 1$  and  $2$ ; the optimum  $Q^*$  values for the given  $(n, m)$  combinations were utilized. In all cases, as expected, the estimators  $\hat{Y}_{2c}$  and  $\hat{Y}_{2c}^*$  in the wor design are more efficient than  $\hat{Y}_{2R}^*$  in Raj's design. As  $n$  increases for a given  $\rho_z$ , the efficiency gain using the wor strategy increases. Finally, we note that Raj realized efficiency gains compared with no matching only when  $\rho_z > 0.5$  whereas efficiency gains always resulted using  $\hat{Y}_{2c}$  for any  $\rho_z$  in the wor situation.

#### ACKNOWLEDGEMENT

We wish to thank Professor J.N.K. Rao for suggestions and comments made during the course of this study. The referee's helpful comments are also much appreciated.

#### REFERENCES

- [1] CHAUDHURI, A. (1980), "On Optimal and Related Strategies for Sampling on Two Occasions with Varying Probabilities", Preprint, Indian Statistical Institute.
- [2] CHOTAI, J. (1974), "A Note on the Rao-Hartley-Cochran Method for PPS Sampling Over Two Occasions", Sankhya, 36, C, 173-180.

- [3] CHOUDHRY, G.H. (1981), "Construction of Working Probabilities and Joint Selection Probabilities for Fellegi's PPS Sampling Scheme", Survey Methodology, 7, No.1, 93-108.
- [4] FELLEGI, I.P. (1963), "Sampling with and Without Replacement: Rotating and Non-rotating Samples", J. Amer. Stat. Ass., 58, 183-201.
- [5] GHANGURDE, P.D. and Rao, J.N.K. (1969), "Some Results on Sampling Over Two Occasions", Sankhya, 31, A, 463-472,
- [6] MURTHY, M.N. (1967), "Sampling Theory and Methods", Calcutta: Statistical Publishing Society.
- [7] RAJ, D. (1965), "On Sampling Over Two Occasions with Propability Proportional to Size", Ann. Math. Statist., 36, 327-330.
- [8] RAO, J.N.K., HARTLEY, H.O. and COCHRAN, W.G. (1962), "On a Simple Procedure of Unequal Probability Sampling without Replacement", J.R. Statist. Soc. B., 24, No. 2, 482-491.



TABLE 1

Efficiencies of composite wor estimators relative to ppswor estimator

Population	N	(n,m)	$\rho_z$	$Q_{opt}$	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}_{2c})}$	$Q_{opt}^*$	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}_{2c}^*)}$
Murthy set 1	17	(2,1)	0.940	0.443	1.337	0.640	1.476
		(3,2)		0.456	1.230	0.738	1.367
		(3,1)		0.419	1.524	0.545	1.656
		(4,3)		0.462	1.188	0.793	1.309
Murthy set 2	16	(2,1)	0.867	0.377	1.205	0.615	1.364
		(3,2)		0.402	1.151	0.727	1.296
		(3,1)		0.336	1.282	0.499	1.459
		(4,3)		0.431	1.191	0.786	1.255
Acreages	14	(2,1)	0.546	0.181	1.083	0.466	0.927
		(3,2)		0.215	1.070	0.646	0.927
		(3,1)		0.137	1.092	0.295	0.933
		(4,3)		0.279	1.117	0.736	0.931
Data set 1	15	(2,1)	0.392	0.113	1.019	0.506	1.013
		(3,2)		0.140	1.017	0.670	1.013
		(3,1)		0.082	1.020	0.340	1.013
		(4,3)		0.330	1.170	0.752	1.012
Data set 2	16	(2,1)	0.213	0.061	1.007	0.451	0.898
		(3,2)		0.078	1.007	0.636	0.896
		(3,1)		0.042	1.007	0.280	0.908
		(4,3)		0.285	1.142	0.730	0.901

TABLE 2

Efficiencies of composite wor estimators relative to Raj's composite estimator

Population	N	(m,m)	$\rho_z$	$\text{Var}(\hat{Y}_{2R}^*)/\text{Var}(\hat{Y}_{2c})$	$\text{Var}(\hat{Y}_{2R}^*)/\text{Var}(\hat{Y}_{2c}^*)$
Murthy set 1	17	(2,1)	0.940	1.038	1.146
		(3,2)		1.127	1.252
		(3,1)		1.215	1.320
		(4,3)		1.248	1.375
Murthy set 2	16	(2,1)	0.867	1.001	1.133
		(3,2)		1.106	1.246
		(3,1)		1.130	1.287
		(4,3)		1.309	1.380
Acreages	14	(2,1)	0.546	1.244	1.065
		(3,2)		1.330	1.151
		(3,1)		1.351	1.154
		(4,3)		1.507	1.257
Data set 1	15	(2,1)	0.392	1.095	1.089
		(3,2)		1.197	1.192
		(3,1)		1.200	1.192
		(4,3)		1.522	1.317
Data set 2	16	(2,1)	0.213	1.197	1.068
		(3,2)		1.293	1.152
		(3,1)		1.280	1.154
		(4,3)		1.589	1.254

# SURVEY METHODOLOGY

1982

Vol. 8

Nos. 1 & 2

A Journal produced by Methodology Staff, Statistics Canada

## CONTENTS

The Role of the Questionnaire in Survey Design R. PLATEK and D. ROYCE .....	1
Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey J.D. DREW, M.P. SINGH and G.H. CHOUDHRY .....	17
Characteristics of Respondent and Non-Respondent Households in the Canadian Labour Force Survey ELIZABETH CLAYTON PAUL and MURRAY LAWES .....	48
Rotation Group Bias in the LFS Estimates P.D. GHANGURDE .....	86
Computerization of Complex Survey Estimates M.A. HIDIROGLOU .....	112



















TABLE DES MATIÈRES

1	Importance du questionnaire dans le plan de sondage R. PLATEK, D. ROYCE .....
19	Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada J.D. DREW, M.P. SINGH, G.H. CHOUDHRY .....
33	Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active du Canada ELIZABETH CLAYTON PAUL, MURRAY LAWES .....
94	Le biais de renouvellement de l'échantillon dans les estimations de l'EPA P.D. GHANGURDE .....
112	Information du calcul d'estimation pour les enquêtes complexes M.A. HIDIRGLOU .....

TABEAU 2

Efficacité des estimateurs composites des plans d'échantillonnage SR par rapport à celle de l'estimateur composite de Raj

Population	N	(n,m)	$P_z$	$\text{Var}(\hat{\psi}_{2R}^*)/\text{Var}(\hat{\psi}_{2C}^*)$	$\text{Var}(\hat{\psi}_{2R}^*)$	$\text{Var}(\hat{\psi}_{2C}^*)$
Murthy	17	(2,1)	0.940	1.038	1.146	1.252
Groupe 1		(3,2)		1.127	1.320	1.252
		(3,1)		1.215	1.320	1.252
		(4,3)		1.248	1.375	1.375
Murthy	16	(2,1)	0.867	1.001	1.133	1.246
Groupe 2		(3,2)		1.106	1.287	1.246
		(3,1)		1.130	1.287	1.287
		(4,3)		1.309	1.380	1.380
Nombre d'acres	14	(2,1)	0.546	1.244	1.065	1.151
		(3,2)		1.330	1.151	1.151
		(3,1)		1.351	1.154	1.154
		(4,3)		1.507	1.257	1.257
Ensemble de données 1	15	(2,1)	0.392	1.095	1.089	1.192
		(3,2)		1.197	1.192	1.192
		(3,1)		1.200	1.192	1.192
		(4,3)		1.522	1.317	1.317
Ensemble de données 2	16	(2,1)	0.213	1.197	1.068	1.152
		(3,2)		1.293	1.152	1.152
		(3,1)		1.280	1.154	1.154
		(4,3)		1.589	1.254	1.254

TABLÉAU 1

Efficacité des estimateurs composites des plans d'échantillonnage SR par rapport à celle de l'estimateur du plan d'échantillonnage avec PPTS

Population	N	(n,m)	$p_z$	$Q_{opt}$	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}_{2c})}$	$Q_{opt}^*$	$\frac{Var(\hat{Y}_2)}{Var(\hat{Y}_{2c}^*)}$
Murthy	17	(2,1)	0.940	0.443	1.337	0.640	1.476
Groupe 1		(3,2)	0.456	1.230	0.738	1.367	
		(3,1)	0.419	1.524	0.545	1.656	
		(4,3)	0.462	1.188	0.793	1.309	
Murthy	16	(2,1)	0.867	0.377	1.205	0.615	1.364
Groupe 2		(3,2)	0.402	1.151	0.727	1.296	
		(3,1)	0.336	1.282	0.499	1.459	
		(4,3)	0.431	1.191	0.786	1.255	
Nombre d'acres	14	(2,1)	0.546	0.181	1.083	0.466	0.927
		(3,2)	0.215	1.070	0.646	0.927	
		(3,1)	0.137	1.092	0.295	0.933	
		(4,3)	0.279	1.117	0.736	0.931	
Ensemble de données 1	15	(2,1)	0.392	0.113	1.019	0.506	1.013
		(3,2)	0.140	1.017	0.670	1.013	
		(3,1)	0.082	1.020	0.340	1.013	
		(4,3)	0.330	1.170	0.752	1.012	
Ensemble de données 2	16	(2,1)	0.213	0.061	1.007	0.451	0.898
		(3,2)	0.078	1.007	0.636	0.896	
		(3,1)	0.042	1.007	0.280	0.908	
		(4,3)	0.285	1.142	0.730	0.901	

- [4] FELLEGI, I.P. (1963). "Sampling With and Without Replacement: Rotating and Non-rotating Samples", J. Amer. Stat. Ass. 58, 183-201.
- [5] GHANGURDE, P.D. et RAO, J.N.K. (1969). "Some Results on Sampling Over Two Occasions", Sankhya 31, A, 463-472.
- [6] MURTHY, M.N. (1967). Sampling Theory and Methods. Calcutta : Statistical Publishing Society.
- [7] RAJ, D. (1965). "On Sampling Over Two Occasions With Probability Proportional to Size", Ann. Math. Statist. 36, 327-330.
- [8] RAO, J.N.K., HARTLEY, H.O. et COCHRAN, W.G. (1962). "On a Simple Procedure of Unequal Probability Sampling Without Replacement", J.R. Statist. Soc. B. 24, no. 2, 482-491.

Le tableau 2 indique l'efficacité relative de  $\hat{V}_{2c}^*$  et  $\hat{V}_{2c}$  dans un plan d'échantillonnage avec PPTS par rapport à l'estimateur  $\hat{V}_{2R}^*$  utilisé par Raj (1965) dans son plan d'échantillonnage avec PPTAR décrit plus haut. Pour que les comparaisons soient plus valables, on n'a pas supposé que  $V_{pdt}(Y_t)$  était égal aux deux reprises  $t = 1$  et  $t = 2$ ; on s'est servi des valeurs optimales de  $Q^*$  pour chaque couple  $(n, m)$ . Dans tous les cas, tel que prévu, les estimateurs  $\hat{V}_{2c}^*$  et  $\hat{V}_{2c}$  dans le plan d'échantillonnage SR s'avèrent plus efficaces que  $\hat{V}_{2R}^*$  dans le plan de Raj. Pour une valeur donnée de  $p_z$ , plus  $n$  est grand, plus le gain d'efficacité entraîné par l'échantillonnage SR est élevé. Enfin, on constate que Raj a réalisé des gains d'efficacité par rapport aux échantillons sans appariement seulement quand  $p_z > 0.5$ , alors que qu'on réalise tous jours des gains d'efficacité si l'on utilise  $\hat{V}_{2c}$  dans l'échantillonnage SR, quelle que soit la valeur de  $p_z$ .

## REMERCIEMENTS

Nous remercions le Professeur J.N.K. Rao pour ses suggestions et ses observations au cours de cette étude, ainsi que l'arbitre pour ses remarques pertinentes.

## BIBLIOGRAPHIE

- [1] CHAUDHURI, A. (1980). "On Optimal and Related Strategies for Sampling on Two Occasions with Varying Probabilities", préirage, Indian Statistical Institute.

- [2] CHOTAI, J. (1974). "A Note on the Rao-Hartley-Cochran Method for PPS Sampling Over Two Occasions", Sankhya 36, C, 173-180.

- [3] CHODHURY, G.H. (1981). "Construction of Working Probabilities and Joint Selection Probabilities for Fellegi's PPS Sampling Scheme", Techniques d'enquête 7, no. 1, 93-108.



0.940 et 0.213. Le tableau 1 indique également les valeurs optimales de  $Q$  et  $Q^*$ . On prévoit effectuer une étude ultérieure qui portera sur l'efficacité de  $\hat{V}_{2c}$  et  $\hat{V}_{2c}^*$  lorsque les valeurs choisies pour  $Q$  et  $Q^*$  ne sont pas optimales.

Les conclusions suivantes se dégagent de ces données empiriques: 1) la valeur optimale de  $Q$  est généralement plus grande quand  $p_z$  est grand et que la valeur optimale de  $Q$  diminue, que l'estimateur soit  $\hat{V}_{2c}$  ou  $\hat{V}_{2c}^*$ . 2) La valeur optimale de  $Q$  pour  $\hat{V}_{2c}^*$  dépasse toujours celle de  $Q$  pour  $\hat{V}_{2c}$ . 3) Lorsque  $p_z$  diminue, l'efficacité de  $\hat{V}_{2c}$  par rapport à  $\hat{V}_2$  diminue aussi (tel que prévu) et se rapproche de l'unité comme limite inférieure quand la valeur de  $Q$  est optimale. Par contre, ce genre de phénomène ne s'observe pas clairement dans le cas de  $\hat{V}_{2c}$  puisque  $\text{Var}(\hat{V}_{2c}^*)$  n'est pas une fonction monotone de  $p_z$ . Pour de petites valeurs de  $p_z$ , on enregistre de faibles gains et de faibles pertes d'efficacité par rapport à  $\hat{V}_2$ . 4)  $\hat{V}_{2c}^*$  est plus efficace que  $\hat{V}_{2c}$  quand la valeur de  $p_z$  est élevée, alors que  $\hat{V}_{2c}$  est plus efficace que  $\hat{V}_{2c}^*$  quand la valeur de  $p_z$  est basse. 5) Si  $\lambda = m/n$  est faible, comme dans le plan (3,1), on peut réaliser d'importants gains d'efficacité si l'on utilise  $\hat{V}_{2c}^*$  au lieu de  $\hat{V}_2$  quand la valeur de  $p_z$  est élevée. Quand  $p_z$  est faible, le plan (4,3) conduit aux meilleurs gains d'efficacité si l'on se sert de  $\hat{V}_{2c}$ ; les résultats des trois autres plans d'échantillonnage sont à peu près les mêmes.

Il convient de noter que même s'il existe une bonne corrélation entre les valeurs de  $y_{1i}$  et  $y_{2i}$ , l'estimateur composite  $\hat{V}_{2c}^*$  peut toujours causer des pertes d'efficacité en comparaison de l'estimateur  $\hat{V}_2$  calculée à partir de données recueillies à une seule reprise au moyen de l'échantillonnage avec PPT. Le facteur critique est la corrélation  $p_z$  entre les valeurs de  $z_{1i}$  et  $z_{2i}$ . La meilleure solution est d'utiliser  $\hat{V}_{2c}$  avec la valeur optimale de  $Q$  puisque  $\hat{V}_{2c} = \hat{V}_2$  quand  $Q = 0$ .

PPT. Comme on ne dispose pas d'une forme fermée de  $\text{Var}(\hat{Y}_{2C})$  et de  $\text{Var}(\hat{Y}_{2C}^*)$

pour permettre des comparaisons analytiques, on a utilisé de petites populations pour obtenir des valeurs de variables et mesurer les contrastes entre les estimateurs. (Les populations étudiées devaient être de petite taille, comme celles qu'on trouve souvent dans les échantillons stratifiés, puisque l'effet de l'échantillonnage avec ou sans remise se voit seulement quand les fractions de sondage sont non négligeables.) Quatre plans d'échantillonnage

avec renouvellement ont été appliqués à chaque population :  $(n, m) = (2, 1)$ ,  $(3, 2)$ ,  $(3, 1)$  et  $(4, 3)$ , où  $n$  est le nombre d'unités de l'échantillon à chaque reprise et  $m$  est le nombre d'unités incluses dans l'échantillon à la première reprise. Deux populations proviennent de l'ouvrage de Murthy

(1967), qui a divisé une seule population de 34 villages en deux populations de 16 et de 17 villages (une unité extrême a été exclue). La caractéristique  $x$  qui sert à mesurer la taille est le nombre d'acres cultivés en 1961;  $Y_1$  et  $Y_2$  sont le nombre d'acres de blé en 1963 et en 1964 respectivement. Une troisième population est un ensemble de 14 fermes dans la province de la Saskatchewan où  $x$  est le nombre d'acres des fermes en 1980 et  $Y_1$  et  $Y_2$  sont le

nombre d'acresensemencés en 1980 et en 1981 respectivement. On analyse aussi deux autres ensembles de données recueillies parmi une population de 15 unités et une autre de 16 unités.

Le tableau 1 montre l'efficacité relative de  $\hat{Y}_{2C}$  et  $\hat{Y}_{2C}^*$  par rapport à  $\hat{Y}_2$  pour chacune des cinq populations et les quatre plans d'échantillonnage. Un paramètre très important dans chaque comparaison est la corrélation  $\rho_Z$  entre  $Z_1 = Y_1/P_1$  et  $Z_2 = Y_2/P_2$ .

$$\rho_Z = \frac{\sum_{i=1}^N P_{1i} Z_{1i} Z_{2i} - Y_1 Y_2}{\sqrt{\left( \sum_{i=1}^N P_{1i} Z_{1i}^2 - Y_1^2 \right) \left( \sum_{i=1}^N P_{2i} Z_{2i}^2 - Y_2^2 \right)}}$$

Les valeurs de  $\rho_Z$  calculées pour les populations étudiées varient entre

$$\text{Var}(\hat{Y}_2') = N^2 \left[ \left( \frac{1}{1} - \frac{n}{1} \right) (S_2^2 - 2S_{12}) + \left( \frac{m}{1} - \frac{1}{1} \right) S_2^2 \right]$$

où, par exemple :

$$S_{12}^2 = \frac{1}{N} \sum_{i=1}^N (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2) / (N-1).$$

Cette expression concorde avec l'équation (1) dans le cas où  $p_i = 1/N$  et  $P(i, j) = n(n-1)/(N(N-1))$  ( $i \neq j$ ).

De plus, dans l'échantillonnage aléatoire simple,  $\hat{Y}_2 = N\bar{Y}_2$  (où  $\bar{Y}_2$  est la moyenne de l'échantillon calculée à partir de toutes les  $n$  unités de l'échantillon constitué à la deuxième reprise) et la variance de  $\hat{Y}_2$  est

$$\text{Var}(\hat{Y}_2) = N(N-n)S_2^2.$$

Une évaluation de  $\text{Var}(\hat{Y}_2)$  au moyen de l'expression (2) produit le même résultat. Enfin, on peut obtenir l'expression suivante soit par une évaluation directe ou à partir de l'équation (3) :

$$\text{Cov}(\hat{Y}_1', \hat{Y}_2') = -NS_2^2$$

De façon semblable, on peut calculer  $\text{Var}(\hat{Y}_{2c}^*)$ .

### 3. EXEMPLES QUANTITATIFS

On compare ici l'efficacité des estimateurs composites  $\hat{Y}_{2c}$  et  $\hat{Y}_{2c}^*$ , auxquels correspond une valeur optimale,  $Q$  et  $Q^*$  respectivement, qui réduit au minimum la variance de l'estimation, et celle de l'estimateur  $\hat{Y}_2$  calculé à partir des renseignements courants recueillis auprès d'un échantillon sélectionné avec

$$\text{Var}(\hat{Y}_{2C}) = [\text{Var}(\hat{Y}_1^*) \cdot \text{Var}(\hat{Y}_2) - (\text{Cov}(\hat{Y}_1^*, \hat{Y}_2))^2] / [\text{Var}(\hat{Y}_1^*) + \text{Var}(\hat{Y}_2) - 2 \text{Cov}(\hat{Y}_1^*, \hat{Y}_2)] \cdot$$

Un autre estimateur composite  $\hat{Y}_{2C}^*$  de  $Y_2$  est

$$\hat{Y}_{2C}^* = Q^* \hat{Y}_1^* + (1 - Q^*) \hat{Y}_{2U},$$

où

$$\hat{Y}_{2U} = \frac{\sum_{i=1}^{n+u} Y_{2i}}{N} = \frac{\sum_{i=1}^n Y_{2i}}{N} + \frac{\sum_{i=n+1}^{n+u} Y_{2i}}{N} = \hat{Y}_{2C} + \frac{\sum_{i=n+1}^{n+u} Y_{2i}}{N} - \frac{\sum_{i=1}^n Y_{2i}}{N}.$$

Pour calculer la variance de  $\hat{Y}_{2C}^*$ , on utilise la formule (1) et les expressions suivantes :

$$\text{Var}(\hat{Y}_{2U}) = \frac{1}{N} \sum_{i=1}^n P_{(i,j)} Y_{2i}^2 + \frac{1}{N} \sum_{i=n+1}^{n+u} P_{(i,j)} Y_{2i}^2 - \frac{1}{N} \sum_{i=1}^n P_{(i,j)} Y_{2i}^2 - \frac{1}{N} \sum_{i=n+1}^{n+u} P_{(i,j)} Y_{2i}^2,$$

$$\text{Cov}(\hat{Y}_1^*, \hat{Y}_{2U}) = \frac{1}{N} \sum_{i=1}^n P_{(i,j)} Y_{1i} Y_{2i} + \frac{1}{N} \sum_{i=n+1}^{n+u} P_{(i,j)} Y_{1i} Y_{2i} - \frac{1}{N} \sum_{i=1}^n P_{(i,j)} Y_{1i} Y_{2i} - \frac{1}{N} \sum_{i=n+1}^{n+u} P_{(i,j)} Y_{1i} Y_{2i},$$

$$+ \frac{1}{N} \sum_{i=1}^n P_{(i,j)} Y_{1i}^2 + \frac{1}{N} \sum_{i=n+1}^{n+u} P_{(i,j)} Y_{1i}^2 - \frac{1}{N} \sum_{i=1}^n P_{(i,j)} Y_{1i}^2 - \frac{1}{N} \sum_{i=n+1}^{n+u} P_{(i,j)} Y_{1i}^2.$$

## 2.3 Cas spécial

Pour vérifier les résultats des calculs précédents, prenons le cas de l'échantillonnage aléatoire simple sans remise.

Ainsi,  $\hat{Y}_1^* = N(\bar{Y}_1 + (\bar{Y}_{2m} - \bar{Y}_{1m}))$ , où  $\bar{Y}_1$  et  $\bar{Y}_{1m}$  sont les moyennes de l'échantillon calculées respectivement pour toutes les unités échantillonnées et pour toutes les unités apparues à l'échantillon précédent et  $\bar{Y}_{2m}$  est la moyenne de l'échantillon calculée pour toutes les unités apparues de l'échantillon courant. Une évaluation directe conduit au résultat :

Si l'on substitue les trois expressions présentées ci-dessus dans l'équation

de  $\text{Var}(\hat{Y}_2)$ , on obtient la formule :

$$\text{Var}(\hat{Y}_2) = \frac{1}{N} p_i \left[ \frac{z_{2i}^2}{n} + (z_{2i} - z_{1i})^2 \left( \frac{m}{1} - \frac{n}{1} \right) \right] + \frac{1}{2} \sum_{i \neq j} \left[ \frac{p(i, j, es)}{n} z_{1i} z_{1j} + \frac{p(i, j, es)}{2} (z_{2i} - z_{1i})(z_{2j} - z_{1j}) + \frac{nm}{2p(i, j, es)} z_{1i} (z_{2j} - z_{1j}) \right] - Y_2^2 \quad (1)$$

En outre,

$$\text{Var}(\hat{Y}_2) = \frac{1}{2} \sum_{i \neq j} p_i z_{2i}^2 + \frac{nm}{2} \sum_{i \neq j} p(i, j, es) z_{2i} z_{2j} - Y_2^2 \quad (2)$$

où  $s^*$  est l'ensemble des  $n$  unités observées à la deuxième reprise et

$$\text{Cov}(\hat{Y}_2, \hat{Y}_2) = \sum_{i \neq j} \left[ \frac{p(i, es, j, es)}{n^2} z_{1i} z_{2j} + \frac{p(i, es, j, es)}{nm} (z_{2i} - z_{1i}) z_{2j} \right] + \frac{1}{2} \sum_{i \neq j} p_i z_{2i} (z_{2i} - z_{1i}) - Y_2^2 \quad (3)$$

Si l'on combine les expressions (1), (2) et (3), on obtient  $\text{Var}(\hat{Y}_{2c})$ .

La valeur optimale du poids  $Q$  qui réduit au minimum  $\text{Var}(\hat{Y}_{2c})$  est

$$Q_{\text{opt}} = [\text{Var}(\hat{Y}_2) - \text{Cov}(\hat{Y}_1, \hat{Y}_2)] / [\text{Var}(\hat{Y}_1) + \text{Var}(\hat{Y}_2) - 2 \text{Cov}(\hat{Y}_1, \hat{Y}_2)] \cdot$$

La variance minimum ainsi obtenue est



$$p_j(k) \times \frac{\dots \times \frac{r-1}{k-1} p_i(k) - \sum_{\ell=r+1}^{\ell} p_i(k)}{p_j(k)}$$

or

$$\text{Var}(\hat{y}^2) = \frac{1}{N} \text{Var} \left[ \sum_{i=1}^N a_i y_i / p_i \right] + \frac{1}{N} \text{Var} \left[ \sum_{i=1}^N a_i (y_i - y_i) / p_i \right] + \frac{mn}{2} \text{Cov} \left[ \sum_{i=1}^N a_i y_i / p_i, \sum_{i=u+1}^N a_i (y_i - y_i) / p_i \right]$$

A l'aide des propriétés des variables nominales  $a_i$  présentées plus haut, on peut démontrer que :

$$\frac{1}{N} \text{Var} \left[ \sum_{i=1}^N a_i y_i / p_i \right] = \frac{1}{N} \sum_{i=1}^N p_i z_i^2 + \frac{1}{2} \sum_{i \neq j} p_i p_j z_i z_j - y_i^2$$

$$\frac{1}{N} \text{Var} \left[ \sum_{i=u+1}^N a_i (y_i - y_i) / p_i \right] = \frac{1}{N} \sum_{i=1}^N p_i (z_i - z_i)^2$$

$$+ \frac{1}{2} \sum_{i \neq j} p_i p_j (z_i - z_i)(z_j - z_j) - (y_i - y_i)^2$$

où  $z_{tj} = y_{tj} / p_j$ ,  $t=1, 2$  et  $n-u=m$ .

En outre,

$$\frac{1}{mn} \text{Cov} \left[ \sum_{i=1}^N a_i y_i / p_i, \sum_{i=u+1}^N a_i (y_i - y_i) / p_i \right]$$

$$= \frac{1}{N} \sum_{i=1}^N p_i z_i (z_i - z_i) + \frac{1}{mn} \sum_{i \neq j} p_i p_j (z_i - z_i)(z_j - z_j) - (y_i - y_i) \cdot$$

est le résultat des propriétés suivantes de la variable nominale  $r_i$ :

$$\begin{aligned} \text{Var}(r_i) &= p_i (1 - p_i), \quad (i = 1, 2, \dots, N, r = 1, 2, \dots, n+u), \\ \text{Cov}(r_i, r_t) &= -p_i p_t, \quad (i \neq t), \\ \text{Cov}(r_i, r_j) &= -p_i p_j, \quad (i \neq j), \\ \text{Cov}(r_i, r_j) &= E(r_i \cdot r_j) - p_i p_j, \quad \text{autrement,} \end{aligned}$$

où  $E(\cdot)$  correspond à l'espérance mathématique de l'expression entre parenthèses dans le plan d'échantillonnage probabiliste. Or  $E(r_i, r_j) = P(r_i = 1, r_j = 1) = P(r_i = 1, r_j = 1)$ , où  $P(\cdot)$  représente une probabilité.

Soit  $\sum_{i,j}^{(k-2)}$  la sommation pour toutes les valeurs possibles des  $(k-2)$ -uples formées de différentes unités  $\{i_1, i_2, \dots, i_{r-1}, i_r, i_{r+1}, \dots, i_{k-2}, i_{k-1}\}$  incluses dans l'échantillon lors des  $k$  premiers tirages à partir des  $N-2$  unités de l'ensemble  $\{1, 2, \dots, i-1, i+1, \dots, j-1, j+1, \dots, N\}$ , où la  $i$ ème unité est sélectionnée au tirage  $r$  et la  $j$ ème unité au tirage  $k$ . Cette sommation comporte  $(N-2)(N-3) \dots (N-k+1)$  termes.

Comme dans l'analyse de Fellegi (1963), définissons  $\{p_i(\lambda); i=1, 2, \dots, N\}$  comme étant l'ensemble des "probabilités de travail" pour la sélection d'une unité au tirage  $\lambda$ ,  $\lambda = 1, 2, \dots, n+u$ . Aux tirages  $k$  et  $r$ , où  $k > r$ :

$$\begin{aligned} E(r_i \cdot r_j) &= \sum_{i,j}^{(k-2)} p_i^{i_1} p_j^{j_1} (1) \frac{1 - p_i^{i_1} (2)}{p_i^{i_1} (2)} \dots \frac{1 - p_i^{i_1} (r-1)}{p_i^{i_1} (r-1)} \\ &\quad \times \frac{p_i^{i_1} (r)}{p_i^{i_1} (r+1)} \times \frac{1 - p_i^{i_1} (r)}{1 - p_i^{i_1} (r+1)} \times \frac{1 - p_i^{i_1} (r)}{1 - p_i^{i_1} (r+1)} \end{aligned}$$

une méthode optimale. Ce résultat offre une raison de plus d'utiliser la méthode de Fellegi.

## 2.2 Méthode d'estimation

Dans cette section, on propose des estimateurs composites de  $Y_2$ , la valeur globale courante, et on évalue leur variance à l'aide d'une variable nominale.

Posons que  $r_{ai} = 1$  si l'unité  $i$ ,  $i = 1, 2, \dots, N$  est sélectionnée au tirage  $r$ ,  $r = 1, 2, \dots, n+u$  et que  $r_{ai} = 0$  autrement. Comme l'espérance mathématique de  $r_{ai}$  est égale à  $p_i$ , un estimateur non biaisé de la valeur globale  $Y_1$  de la population calculée la première fois s'écrit :

$$\hat{Y}_1 = \frac{1}{n} \sum_{r=1}^n \sum_{i=1}^N r_{ai} Y_{1i} / p_i.$$

$$\text{Donc, } \hat{Y}'_2 = \hat{Y}_1 + \frac{1}{n} \sum_{r=u+1}^m \sum_{i=1}^N r_{ai} (Y_{2i} - Y_{1i}) / p_i$$

est un estimateur non biaisé de la valeur globale  $Y_2$  obtenue la deuxième fois. Un estimateur non biaisé de  $Y_2$  qui est une fonction des observations recueillies à la deuxième reprise est

$$\hat{Y}_2 = \frac{1}{n+u} \sum_{r=1}^n \sum_{i=1}^N r_{ai} Y_{2i} / p_i$$

Un estimateur composite de  $Y_2$  correspond à la moyenne pondérée

$$\hat{Y}_{2c} = q \hat{Y}'_2 + (1-q) \hat{Y}_2,$$

$$\text{où } 0 \leq q \leq 1.$$

La variance de  $\hat{Y}_{2c}$ ,  $\text{Var}(\hat{Y}_{2c}) = q^2 \text{Var}(\hat{Y}'_2) + (1-q)^2 \text{Var}(\hat{Y}_2) + 2q(1-q) \text{Cov}(\hat{Y}'_2, \hat{Y}_2)$ ,

L'estimateur composite  $\hat{V}_{2c}^1$  est calculée, Chotai détermine la valeur optimale de  $\lambda$  et compare l'efficacité relative de  $\hat{V}_{2c}^1$  par rapport aux estimateurs optimaux de Changuide et Rao et de Raj. Il s'est avéré que  $\hat{V}_{2c}^1$  est toujours plus efficace que  $\hat{V}_{2R}^*$  et, dans bien des cas, plus efficace que  $\hat{V}_{2G}^1$ . Chotai examine brièvement le cas où  $n/m$  n'est pas un nombre entier. Il convient de noter que, étant donné que  $\lambda$  n'est pas une fonction continue, la valeur optimale de  $\lambda$  devrait être calculée par des méthodes de programmation en nombres entiers. Dans la section suivante, on présente une méthode d'échantillonnage qui conduit souvent à des résultats plus efficaces que ceux obtenus au moyen des techniques d'échantillonnage avec PPSR proposées jusqu'à présent.

## 2. MÉTHODE D'ÉCHANTILLONNAGE

### 2.1 Sélection de l'échantillon

Dans la population de  $N$  unités  $(1, 2, \dots, N)$ , choisissons un échantillon qui contient  $n + u$  unités ( $u < n$ ) tirées une à la fois sans remise selon la méthode de de Fellegi, où la probabilité de sélectionner la  $i$ ème unité à chaque tirage est  $p_i$ ,  $i = 1, 2, \dots, N$ ,  $\sum p_i = 1$ . Lors du premier échantillonnage, les  $n$  premières unités des  $n + u$  unités sont examinées, tandis qu'au deuxième échantillonnage, les  $u$  premières unités sont supprimées de l'échantillon et les  $n$  unités non étudiées la première fois sont incluses. Ainsi,  $m = n - u$  unités sont observées dans les deux cas. Les  $n$  unités étudiées la première fois composent l'échantillon  $s_1$  (où  $s_1 \subset s$ ) et l'ensemble d'unités introduites seulement à la deuxième période représentent l'échantillon  $s_2$ . Soulignons que la méthode de Fellegi garantit que la probabilité de sélection d'une unité  $i$  est la même à chaque tirage et donc au cours des deux périodes. Chaudhuri (1980) a étudié une sous-catégorie d'estimateurs non biaisés de modèles linéaires non homogènes et il a démontré que le plan d'échantillonnage exposé plus haut conduit à

Fellegi étaient beaucoup trop compliqués pour  $n > 2$ . Choudhry (1981) a conçu une méthode itérative pour mettre en oeuvre la technique de Fellegi et il a élaboré un programme informatique qui évalue les probabilités de travail pour  $n \leq 5$ . Même si la convergence est rapide du point de vue du nombre d'itérations, le nombre de calculs augmente par une proportion égale à  $N^n$ . Le programme indique aussi pour le calcul de la variance la probabilité composée que les unités  $i$  et  $j$  soient incluses en même temps dans l'échantillon.

Rao, Hartley et Cochran (1962) ont mis au point la "méthode des groupes aléatoires" pour la sélection d'un échantillon avec PPSR. La population de  $N$  unités est divisée en  $n$  groupes de taille  $N_1, N_2, \dots, N_n$ , où  $\sum N_h = N$ , et un échantillon d'une unité est prélevé indépendamment de chaque groupe avec probabilité proportionnelle aux valeurs de  $p_i$ . Ghangurde et Rao (1969) ont adapté la méthode des groupes aléatoires à l'échantillonnage à deux reprises. Pour simplifier la description, les  $N$  unités sont réparties en  $n$  groupes de taille  $N/n$  (nombre qu'on suppose entier). Dans un premier temps, une unité est tirée de chaque groupe aléatoire, tel qu'il a été mentionné plus haut, pour former un échantillon  $s$  de  $n$  unités. A un deuxième moment, un échantillon aléatoire simple  $s_1$  de  $m$  unités apparues est sélectionné SR parmi les  $n$  unités et un échantillon indépendant  $s_2$  de  $u = n - m$  unités est recueilli à partir de l'ensemble de la population par la même méthode appliquée pour produire l'échantillon  $s$ . Ghangurde et Rao utilisent un estimateur composite de  $\bar{Y}_{2G}$  de  $\bar{Y}_2$  pour réduire au minimum la variance de cette caractéristique à l'aide d'une valeur optimale du poids  $Q$ . Ensuite, la valeur optimale de  $\lambda = m/n$  est déterminée. Ces auteurs font remarquer qu'il serait probablement plus efficace de tirer  $s_2$  parmi les  $N - n$  unités de la population non incluses dans  $s$ .

Chotali (1974) a modifié la méthode de Ghangurde et Rao (G-R) pour l'échantillonnage à la deuxième reprise; les  $n$  unités de  $s$  sont choisies au hasard pour former  $m$  groupes de taille  $n/m$  (nombre qu'on suppose entier). Une unité est tirée de chacun des  $m$  groupes avec probabilité proportionnelle aux valeurs de  $p_i$ , pour constituer un échantillon  $s_1$ . Ensuite, un échantillon  $s_2$  est prélevé à l'aide de la méthode G-R. Une fois que la variance optimale de



Raj (1965) a étudié le plan d'échantillonnage suivant avec PPT (probabilité proportionnelle à la taille) : un premier échantillon  $s$  de taille  $n$  est constitué avec probabilités  $p_i$  proportionnelles aux valeurs de  $x_i$  et avec remise (AR). Dans une deuxième période, un échantillon aléatoire simple  $s_1$  de  $m$  unités est tiré de  $s$  sans remise (SR) et un échantillon indépendant  $s_2$  de  $m$  unités  $u = n - m$  est sélectionné avec PPT à partir de l'ensemble de la population. Les valeurs de  $y_1$  et  $y_2$  sont alors estimées à l'aide des équations :

$$\hat{y}_1 = \sum_{i \in s} y_{1i} / (np_i) \text{ and } \hat{y}_2^* = q^* \hat{y}_2 + (1 - q^*) \hat{y}_2',$$

$$\text{où } \hat{y}_2 = \sum_{i \in s_2} y_{2i} / (np_i), \hat{y}_2' = \hat{y}_1 + \sum_{i \in s_1} (y_{2i} - y_{1i}) / (mp_i),$$

et  $q^*$  est un poids,  $0 \leq q^* \leq 1$ .

La variance de  $\hat{y}_2^*$  a été réduite au minimum à partir de l'hypothèse selon laquelle l'expression

$$v_{pps}(y_t) = \sum_{i=1}^N p_i (y_{ti}/p_i - y_t)^2$$

est identique pour les temps  $t = 1$  et  $t = 2$ .

Les ouvrages statistiques ont prêté une attention considérable au problème de l'échantillonnage à une reprise avec PPTS. Une des difficultés principales est la spécification de méthodes pratiques pour obtenir certaines probabilités données à chaque tirage. Fellegi (1963) a proposé une technique où la probabilité de sélectionner l'unité  $i$  à chacun des  $n$  tirages est égale à  $p_i$  si l'on détermine  $n - 1$  ensembles de "probabilités de travail". Ce procédé est extrêmement important pour le renouvellement d'échantillons, où il est essentiel que l'estimateur habituel de  $Y_2$  pour les échantillons constitués avec PPT ne soit pas biaisé; cet objectif ne sera pas atteint si  $p_i$  n'est pas constant à chacun des  $n$  tirages d'échantillonnage avec renouvellement partiel. Jusqu'à récemment, les calculs nécessaires pour la technique de

## ÉCHANTILLONNAGE À DEUX REPRISSES AVEC PPSR

G.H. Choudhry et Jack E. Graham<sup>1</sup>

On décrit une théorie d'échantillonnage à deux reprises avec probabilités inégales et sans remise. La méthode de Fellgett (1963), où à chaque reprise la même probabilité de sélection est attribuée à une unité donnée, est utilisée pour choisir les unités de l'échantillon de renouvellement. On examine ensuite le développement mathématique de la variance des estimateurs composites de la valeur globale de la population à la deuxième reprise. Des résultats quantitatifs pour des échantillons de petite taille sont présentés et on compare l'efficacité de cette méthode à celle d'une technique différente.

## 1. INTRODUCTION

Dans les enquêtes à caractère répétitif, l'utilisation d'un plan d'échantillonnage avec remise partielle présente certains avantages tant du point de vue de l'efficacité que de la réduction du fardeau du répondant. Essentiellement, après chaque tirage d'un échantillon, une fraction des unités sont supprimées et remplacées par un nouvel sous-échantillon de la population. Un très grand nombre d'ouvrages statistiques examinent les méthodes et les estimateurs que l'on utilise pour l'échantillonnage à deux reprises ou plus avec probabilités égales. Mais les cas où les probabilités de sélection sont inégales présentent encore plus d'importance du point de vue de la pratique. Ainsi, prenons une population finie de  $N$  unités  $\{1, 2, \dots, N\}$  qui est échantillonnée à deux reprises, soit en période 1 (l'échantillon précédent) et en période 2 (l'échantillon courant). Soient  $y_1$  et  $y_2$  les valeurs d'une caractéristique  $y$  de la  $i$ ème unité dans les échantillons 1 et 2 et  $Y_1$  et  $Y_2$  les valeurs globales de la population calculées à partir des échantillons correspondants. Une mesure de taille,  $x_i$ , est connue pour toutes les unités de la population.

<sup>1</sup> G.H. Choudhry, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, et Jack E. Graham, Université Carleton.

FIGURE 2

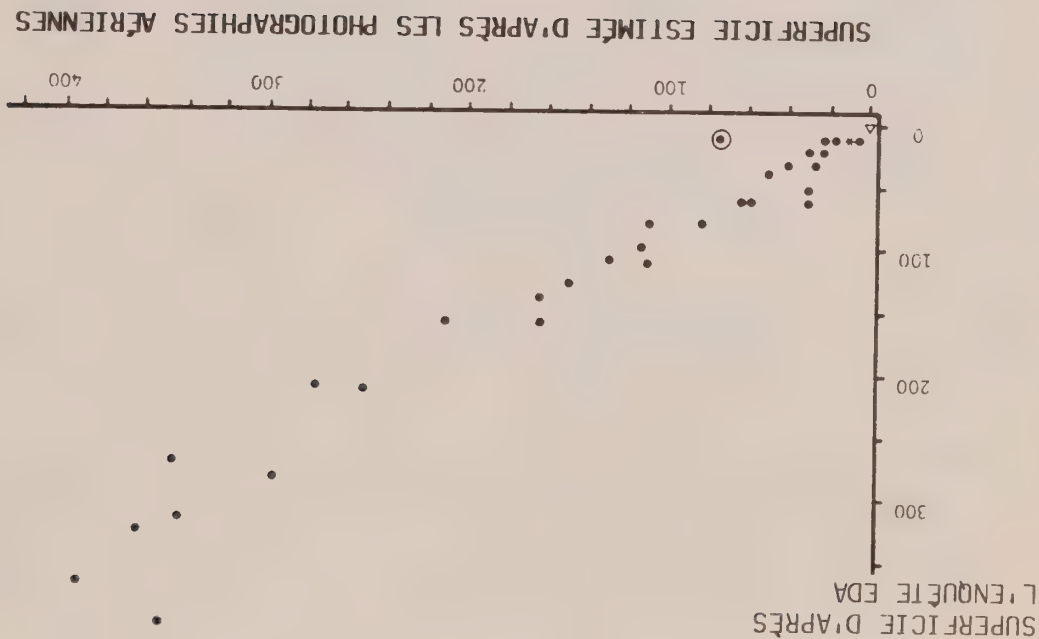
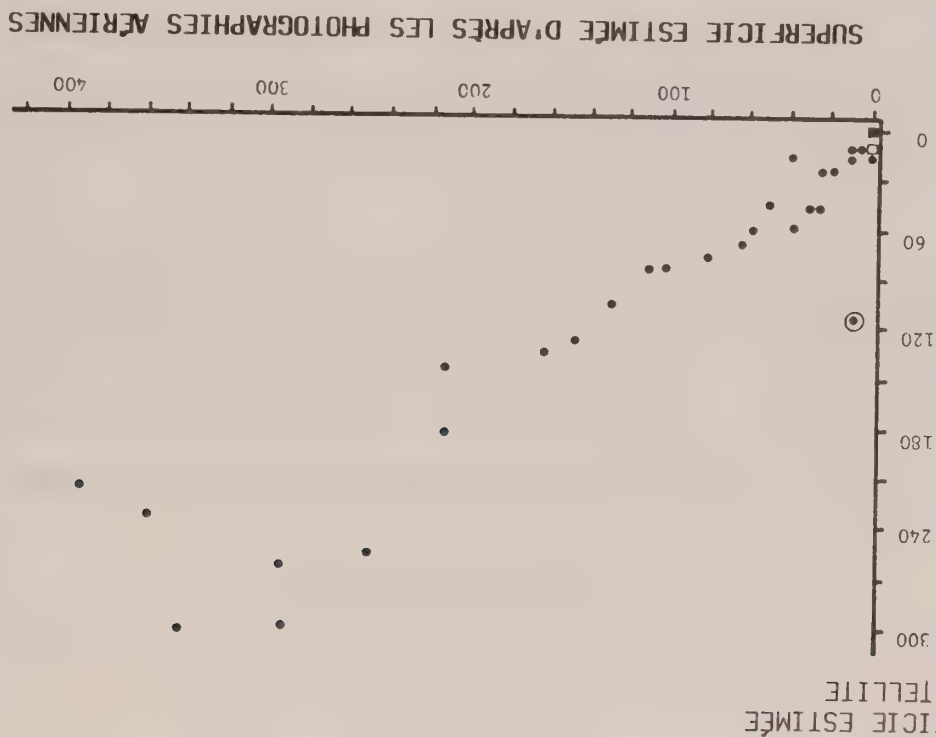


FIGURE 3



FIGURES 2 et 3 - Graphes des superficies estimées, (en acres), respectivement, par l'EDA et par satellite en fonction des superficies estimées d'après les photographies aériennes à basse altitude, pour des segments échantillonnés du Nouveau-Brunswick. Les points encadrés représentent les valeurs aberrantes.

Légende: • = 1 observation, □ = 2 observations, \* = 3 observations, ■ = 11 observations et Δ = 15 observations

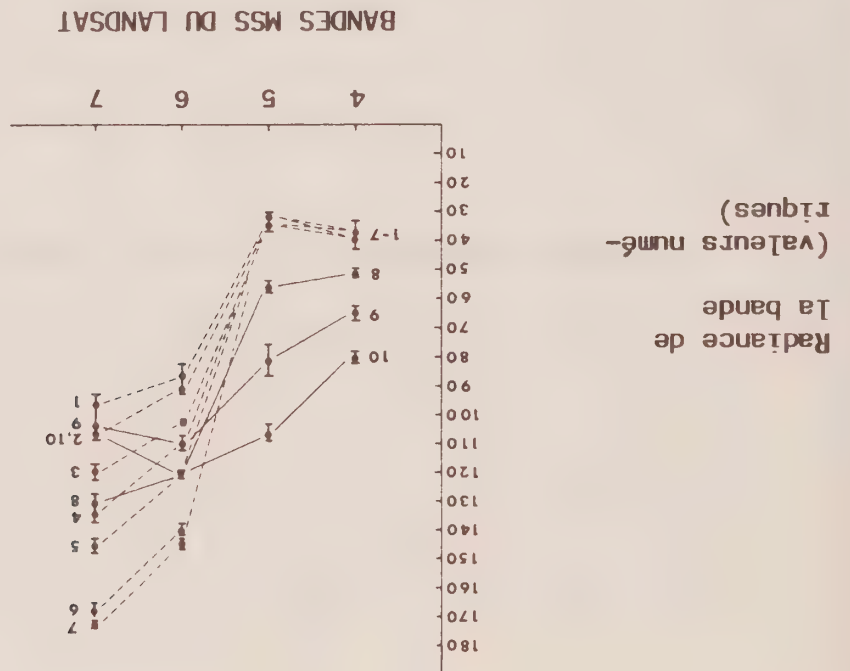


FIGURE 1 COMPARAISON DES DIVERSES BANDES LANDSAT POUR DIFFÉRENTES RÉCOLTES

Radiance des bandes MSS de Landsat (8 août 1975) pour les céréales (1), le sarrasin (2), les pâturages (3), le foin (4), le maïs (5), les pois (6), les pommes de terre (7), le brocoli (8), le sol nu (9), les herbes (10). Les bandes sont numérotées de 4 à 7. L'espacement relatif entre les valeurs numériques dans chaque colonne indique le degré de séparabilité des récoltes pour une bande donnée.

TABEAU I

ESTIMATIONS DE LA SUPERFICIE TOTALE PLANTÉE EN POMMES DE TERRE  
DANS LA RÉGION D'ÉTUDE ET POUR LA PROVINCE DU NOUVEAU-BRUNSWICK<sup>1</sup>

Enquête et/ou	Estimation pour	Estimation pour le	Coefficient de
moyen	la région d'étude	Nouveau-Brunswick	variation (%)
d'estimation	(en acres)	(en acres)	

Satellite			
Données non corrigées	42,354	n.d.	n.d.
Estimation par le quotient	49,504	51,524	5.5
Estimation par régression	49,115	51,119	5.5
Enquête postale	n.d.	50,800	n.d.
EDA à bases multiples	n.d.	53,854	5.2
FORPT	47,203	49,129	4.59
Estimation de Statistique Canada publiée	n.d.	52,000	n.d.
Le 5 septembre			

<sup>1</sup> La région d'essai, regroupant les comtés de Carleton, de Madawaska et le comté Victoria, représente environ 96.08 % des terrains producteurs de pommes de terre au Nouveau-Brunswick.



- [7] Mosher, P.N., R.A. Ryerson et W.M. Strome, 1978, "New Brunswick Potato Area Estimation from Landsat", 12th International Symposium on Remote Sensing of Environment, Manille, Philippines avril 1978, pp. 1415-1419.
- [8] Ryerson, R.A., P. Mosher, V. Wallen et N. Stewart, "Three Tests of Agricultural Remote Sensing in Eastern Canada: Results, Problems and Prospects", Canadian Journal of Remote Sensing, 5(1) 1979, pp. 53-66.
- [9] Ryerson, R.A., P.N. Mosher et Harvie J., 1980, "Potato Area Estimation Using Remote Sensing Methods", CCRS Users Manual 80-2, Énergie, Mines et Ressources, Ottawa, février 1980.

assez précisément la superficie plantée en pommes de terre dans trois importants comtés producteurs de cette culture, au Nouveau-Brunswick. Tout au long du projet décrit dans cette communication, les auteurs ont raffiné les méthodes d'analyse et les procédures sur le terrain, afin d'accroître la précision des estimations, de réduire le fardeau imposé aux répondants et de fournir des données spatiales détaillées qui n'étaient disponibles auparavant que pour les années de recensement.

## 10. REMERCIEMENTS

Les auteurs désirent remercier pour leurs conseils techniques et leurs commentaires rédactionnels Mme N. Chinnappa, de Statistique Canada, et J. Cihlar et F. Ahern du Centre canadien de télédétection.

## 11. BIBLIOGRAPHIE

- [1] Ahern, F.J. et J. Murphy, 1978, "Radiometric Calibration and Correction of Landsat 1, 2, 3, MSS Data", CCRS Research Report 78-4, Énergie, Mines et Ressources, Ottawa, Canada.
- [2] Brown, R.J., Ahern, F.J., Ryerson, R.A., Thomson, K.P.B., Goodenough, D.G., McCormick, J.A. et Teillet, P.M., 1980, "Rapeseed: Operational Monitoring", 6<sup>ème</sup> Symposium canadien sur la télédétection, Halifax, Nouvelle-Écosse, Canada pp. 321-330.
- [3] Cochran, W.G., 1977, Sampling Techniques (3rd Edition), John Wiley and Sons Inc., New York.
- [4] Goodenough, D.G., 1979, "The Image Analysis System (CIAS) at the Canada Centre for Remote Sensing". Canadian Journal of Remote Sensing 5(1), mai 1979, 3-17.
- [5] Guertin, F.G., 1979, Butlin, T.J. and Jones, R.G., 1979, "Correction géométrique des images Landsat au Centre canadien de télédétection", Canadian Journal of Remote Sensing, 5(2), décembre 1979.
- [6] Hanuschak, G., R. Stigman, M. Craig, M. Ozga, R. Luebke, P. Cook, D. Kieweno, and C. Miller, 1979, "Crop - Area Estimates with Landsat Transition from Research and Development to Timely Results", Proceeding of the Fifth Machine Processing of Remote Sensed Data Symposium, Lafayette, Indiana.

En 1980, on a utilisé des données recueillies par satellite pour estimer

## 9. RÉSUMÉ ET CONCLUSIONS

ments qu'au niveau des champs.

recueillies par photographie aérienne était plus forte au niveau des seg- généralement plus grande. La relation entre les données EDA et les données à l'est de la rivière Saint-Jean, où la superficie moyenne des champs était Les déclarations. On a observé davantage de variations dans la région sise rieur des segments d'échantillon, ou bien à la variabilité des erreurs dans teurs comme l'emplacement géographique, la structure des champs à l'inté- ces pourraient être attribuables aux enquêteurs, mais aussi à d'autres fac- déclarations des répondants, selon le secteur d'affectation. Ces différen- zone assignée aux enquêteurs, indiquent une différence possible dans les fonction des superficies estimées d'après les photographies aériennes, par souvent indiquées en multiples de 5 acres. Les graphes des données EDA en obtenues en interrogeant les exploitants agricoles, et ainsi, elles ont été aériennes. Les données de l'EDA sur la superficie des champs ont été autre cause de l'écart entre les données EDA et celles des photographies Une description erronée des limites et de la superficie des champs fut une

satellite grâce aux données recueillies par les agents recenseurs de l'EDA.

lonnés, est encourageante et justifie la correction des estimations par sa- de l'EDA et celle des photographies aériennes, pour les segments échantil- d'après les photographies aériennes. La forte relation entre les données de fournir des superficies correspondant davantage aux superficies estimées projet de démonstration. On croit que ces modifications permettront à l'EDA de superficie qui seront réalisées l'an prochain, lors de la reprise du s'assurer que toutes les fermes soient répertoriées, en vue des estimations Les terrains pour l'EDA de 1981 contenait des instructions spécifiques pour d'échantillon). Pour corriger cette situation, le guide des procédures sur ainsi que des fermes dont les terrains couvraient plus d'un segment omises, on comptait les fermes qui étaient exclues de la base aréolaire,

mes de terre par segment et calculées selon les données recueillies par les agents recenseurs de l'EDA et par satellite, respectivement en fonction des superficies obtenues d'après les photographies aériennes. On n'a pas utilisé tous les segments d'échantillons dans l'analyse. Il n'y avait pas de données par satellite pour 16 segments, en raison de la couverture nuageuse et de la localisation des images. De plus, on n'a pas utilisé 8 segments de l'EDA, car ces segments contenaient des non-répondants à l'enquête et des fermes d'une certaine importance pour lesquelles les agents recenseurs n'avaient pas à recueillir de données sur les champs. Pour les calculs de cette section et ceux de la section 8.2, nous n'avons pas utilisé les deux valeurs aberrantes de ces graphes.

Ces graphes indiquent une très forte relation linéaire entre les données de l'EDA et celles obtenues par les photographies aériennes, mais il en est de même pour les données par satellite comparées aux données obtenues par photographie aérienne au niveau de segment: les coefficients de corrélation<sup>4</sup> sont respectivement<sup>5</sup> de 0.991 et de 0.968. On note une tendance à la sous-estimation des superficies dans le cas des agents recenseurs de l'EDA et dans celui des données obtenues par satellite. Cette tendance est cependant moins marquée pour l'EDA. On a déjà expliqué à la section 7 les causes des écarts dans le cas des superficies calculées par les données satellite. La superficie de certains segments contenant peu ou pas de pommes de terre a été surestimée, à cause du brouillage par d'autres récoltes (certaines données satellite ont été rejetées, car elles confondaient un grand champ de foin parsemé de pierre avec un champ de pommes de terre - on aurait pu éliminer cette erreur en modifiant la classification). Quant à la sous-estimation dans le cas de l'EDA, une des principales causes était que les enquêteurs n'avaient pas énumérés certaines fermes se trouvant dans les segments à cause des procédures à bases multiples (parmi les fermes

---

<sup>4</sup> Les coefficients de corrélation ont été calculés en utilisant la pondération du plan d'échantillonnage.

<sup>5</sup> La représentation graphique des données obtenues par satellite par rapport aux données de l'EDA a été très similaire à celle de la Figure 3, le coefficient de corrélation s'établissant dans ce cas à 0.957.

$$v(\hat{Y}^{reg.}) = v(\hat{Y}) - 2B \text{cov}(\hat{Y}, \hat{X}) + B^2 v(\hat{X}),$$

$$\text{ou } v(\hat{Y}) = \frac{L}{\sum_{h=1}^H n_h(n_h-1)} \sum_{i=1}^L \left[ \frac{M_{hi}}{n_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} - \frac{1}{n_h} \sum_{i=1}^L \frac{M_{hi}}{n_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \right]^2$$

et  $v(\hat{X})$  est calculé en substituant  $x_{hij}$  pour  $y_{hij}$  dans l'expression pour

$$v(\hat{Y}).$$

On peut voir que  $B = \text{cov}(\hat{Y}, \hat{X}) / v(\hat{X})$  est la valeur de  $B$  qui minimise  $v(\hat{Y}^{reg.})$ .

Le tableau I illustre les estimations par le quotient et par régression, ainsi que les estimations de Statistique Canada pour la superficie plantée en pommes de terre au Nouveau-Brunswick. Les estimations par le quotient et par régression, calculées au prorata du niveau provincial, sont très près de la superficie calculée par Statistique Canada, laquelle se chiffre à 52,000 acres. En fait, il y a très peu de différence entre les deux estimations, et elles possèdent toutes deux le même coefficient de variation. Afin de donner une idée du meilleur rendement obtenu grâce à l'estimation par le quotient, la variance de cette estimation ne représentait que le 1/5 de celle de  $\hat{Y}$ , qui est l'estimation théorique de  $Y$  basé sur le plan d'échantillon avec post-stratification (échantillon aréolaire seulement). On peut noter que le coefficient de variation des estimations par le quotient et par régression sont du même ordre que celui de l'estimation à bases multiples obtenu par l'EDA, bien que cette dernière soit basée sur un échantillon plus important.

### 8.3 COMPARAISON DES DONNÉES AU NIVEAU DES SEGMENTS ET DES CHAMPS

Les figures 2 et 3 contiennent les graphes des superficies plantées en pom-



$$\hat{Y}^{quot.} = R \bar{X} = \frac{\bar{X}}{\bar{Y}} \bar{X}, \text{ et}$$

$$\hat{Y}^{reg.} = \hat{Y} + \hat{B} (X - \bar{X}),$$

$$\text{ou } \hat{Y} = \frac{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij}}{\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij}} \text{ est l'estimation théorique de } Y;$$

$\bar{X}$  est l'estimation théorique de  $X$ , qui représente la superficie totale des champs de pommes de terre obtenue par satellite et non corrigée pour la région d'essai.  $\bar{X}$  est obtenu en substituant  $X_{hij}$  pour  $Y_{hij}$  dans la formule pour  $\hat{Y}$ ;

$Y_{hij}$  et  $X_{hij}$  sont les valeurs observées pour le  $j$ -ème segment choisis dans la  $i$ -ème unité (SD) de premier degré choisie pour la strate  $h$ ;

$N_h$  et  $n_h$  sont respectivement le nombre total et le nombre échantillonné de secteur de dénombrement dans la strate  $h$ ,  $h$  prenant les valeurs de 1 à  $L$ ;

$M_{hi}$  et  $m_{hi}$  sont, respectivement, le nombre total et le nombre échantillonné d'unités (segments) de second degré dans le  $i$ -ème SD choisis dans la strate  $h$ ;

$R$  est une estimation de  $Y/X$ ; et, finalement,

$B = \text{cov}(\hat{Y}, \hat{X}) / \text{var}(\hat{X})$  est le coefficient de régression linéaire.

Les estimations de variance de  $\hat{Y}^{quot.}$  et  $\hat{Y}^{reg.}$  sont données par les relations suivantes:

$$v(\hat{Y}^{quot.}) = v(\hat{Y}) - 2R \text{cov}(\hat{Y}, \hat{X}) + R^2 v(\hat{X}), \text{ et}$$

timations au niveau du segment. Comme complètement d'analyse, on a examiné la variation des superficies indiquées lors de l'EDA, pour chaque secteur assigné à un enquêteur.

## 8.2 ESTIMATION DE LA SUPERFICIE PLANTÉE EN POMMES DE TERRE À L'AIDE DES DONNÉES OBTENUES PAR SATELLITE.

Pour les estimations par le quotient et par régression, on a estimé la superficie totale des champs de pommes de terre dans la région d'étude du Nouveau-Brunswick, recourant pour ce faire à des données obtenues par satellite et à des données sur les segments obtenues par photographie aérienne. Les estimations, ainsi que leurs variances, ont été calculées selon le plan d'échantillonnage de l'EDA. Pour le Nouveau-Brunswick, l'EDA est une enquête à base multiple utilisant un échantillon stratifié à deux degrés par répliquat de segments, cet échantillon étant conçu pour donner des estimations précises sur divers caractères au niveau provincial (voir les sections 4 et 6). Comme les strates de l'EDA ne coïncidaient pas avec les limites de la région d'étude, on a eu recours à la technique de la post-stratification afin de calculer les estimations, traitant les secteurs de dénombrement comme des échantillons avec remplacement. On n'a pas incorporé dans l'échantillon les segments pour lesquels les données par satellite étaient manquantes, ni un segment représentant une valeur aberrante (voir la figure 3). Ces estimations ont été basées sur 40 segments. En dernier lieu, comme les agents recenseurs de l'EDA ne recueillaient pas toujours les données sur toutes les fermes se trouvant à l'intérieur du segment (voir la section 8.3), on a utilisé la superficie du segment calculée d'après les photographies aériennes.

Soit "x" la donnée sur les pommes de terre obtenue par satellite et "y" une donnée obtenue par photographie aérienne. Les estimations par le quotient et par régression de Y, qui représente la superficie totale des champs de pommes de terre dans la région d'essai, ont été calculées à l'aide des formules suivantes:

estimations par le quotient et par régression. Quant aux problèmes du brouillage par d'autres récoltes, on les a, du moins pour la plupart, résolus par une légère modification des paramètres de classification. Dans un cas, le brouillage était dû à une forme non-identifiée de repousse naturelle. Le fortement diffusée dans des régions de coupes forestières. Dans d'autres cas, des champs de foin dans lesquels se trouvaient des amoncellements de pierres espacées régulièrement ont présenté une signature spectrale similaire à celle des pommes de terre. Après modification de la classification, les seuls éléments de brouillage qui demeuraient étaient dus à quelques parcelles de trèfles dans un segment et à certaines allées d'un terrain de golf.

Le calcul des superficies est présenté plus en détail dans les paragraphes suivants.

## 8. RÉSULTATS ET ANALYSE DU PROJET D'ÉVALUATION DE LA SUPERFICIE PLANTÉE EN POMMES DE TERRE AU NOUVEAU-BRUNSWICK (1980)

### 8.1 INTRODUCTION

Dans le cadre du projet d'évaluation de la superficie plantée en pommes de terre au Nouveau-Brunswick, réalisé en 1980, on a analysé les données au niveau de la région, des segments et des champs. On a obtenu deux types d'estimation pour la superficie totale des récoltes de pommes de terre dans la région cible de la vallée de la rivière Saint-Jean. Les estimations proviennent des données recueillies par satellite et de données obtenues par photographies aériennes en haute altitude, ces dernières ayant fait l'objet de vérification au sol (Les photographies aériennes ont été obtenues et analysées par le CCT). Les estimations et leurs variances ont ensuite été comparées aux autres estimations obtenues par Statistique Canada lors d'enquêtes réalisées au Nouveau-Brunswick. On a ensuite comparé les superficies plantées en pommes de terre calculées par l'FDA et par satellite à celles qui furent calculées à partir des photographies aériennes (et qui représentaient, dans ce cas-ci, la meilleure approximation des valeurs réelles) afin de déterminer le degré de concordance de ces deux es-

région (près de Saint-André), ainsi que plusieurs autres champs dans un segment situé au sud de la région (près de Hartland).

Après la réception des données satellite, il fut relativement aisé de rap-  
 pelez les frontières de segment, de les superposer, de localiser les champs  
 d'apprentissage et de commencer la classification selon les méthodes déci-  
 tes à la section 7.1. En plus du choix des régions d'apprentissage, on a  
 retenu un autre groupe de champs de grande superficie, ainsi que des sec-  
 teurs de champs connus de terres qui apparaissaient plus rouges  
 sur l'écran que les champs faisant partie de l'ensemble d'apprentissage. À  
 mesure que les résultats de la classification devenaient disponibles, les  
 superficies cultivées étaient enregistrées pour chaque sous-scène de  
 512 x 512 pixels et pour 17 segments de l'EDA.

Les analystes ont rencontré quatre problèmes au cours de la classification;  
 un de ces problèmes avait trait à l'interpolation des récoltes se trouvant  
 sous des nuages épars et sous l'ombre des nuages, alors que les trois  
 autres problèmes avaient trait au brouillage par d'autres récoltes. La  
 méthode d'interpolation pour les champs de pommes de terre se trouvant sous  
 les nuages était relativement simple. On a supposé que la proportion de  
 récoltes sous les nuages était similaire à celle de la région adjacente qui  
 semblait contenir des pommes de terre. Pour déterminer la superficie des  
 champs de pommes de terre sous les nuages (la valeur PC), on a utilisé la  
 formule suivante:

$$PC = \frac{P_M - T_A}{C_A} \times C_A$$

où:  $P_M$  = superficie des champs de pommes de terre mesurée dans la région  
 sans nuage;  $C_A$  = superficie recouverte par les nuages et l'ombre des nua-  
 ges dans la région et  $T_A$  = superficie totale de la région. Les estima-  
 tions totales par satellite ont incorporé ces régions, alors qu'on n'a in-  
 clus aucun segment EDA sous les nuages dans les analyses pour les



en temps réel en raison de la panne de Landsat-II. Un passage de Landsat-II le 17 août a été utilisé pour produire une bande d'ordinateur traitée sur le système DICS et livrée au Centre d'analyse le 22 août, bien avant la date prévue.

La prélocalisation des segments de l'EDA a été réalisée au printemps de 1980 à l'aide de la fonction "cursueur de polygonation" sur le Système d'analyse des images du CCT (Goodenough, 1977). Les frontières des segments de l'EDA ont été fournies par Statistique Canada sur des cartes au 1:50,000<sup>e</sup> et sur des photocopies d'agrandissement de photographies aériennes, remises aux agents recenseurs de l'EDA pour leur énumération. Bien que les limites de certains segments furent faciles à localiser (ruisseaux, lisières de boisés, lacs, etc.), d'autres frontières géographiques ou de recensement, étaient beaucoup plus complexes. Pour les segments dont les limites consistaient en une combinaison de routes principales et (ou) des rivières importantes, on pouvait les localiser, les délimiter et les mettre en mémoire en moins de cinq minutes de travail sur l'image couleurs de 128x128 pixels de 50m agrandie sur un écran de 512x512 lignes. Les segments les plus complexes ont nécessité jusqu'à une heure de travail - la moyenne s'établissant entre 20 et 25 minutes. Une fois localisé, la frontière du segment était mise en mémoire selon des coordonnées de pixel propres au système DICS, de sorte qu'elle pouvait être superposée sur de nouvelles données lorsque venait le temps de localiser les données d'apprentissage et les intrants des estimateurs. Lors de la phase préparatoire, on n'a pu localiser qu'un certain nombre de frontières de segment, car les autres segments chevauchaient deux bandes d'ordinateur DICS ou présentaient d'autres problèmes similaires. On travaillait actuellement sur un logiciel qui permettra de réduire le temps requis pour tout le projet, particulièrement pour la phase de prélocalisation. L'utilisation des photographies originales au lieu des photocopies, prévue pour le volet 1981 du projet (avec 15 nouveaux segments en 1981), devrait aussi permettre de réduire le temps requis.

Une fois les données au sol de 1980 achevinées au centre d'analyse, on procéda au choix des sites potentiels d'apprentissage (selon la taille des champs). On a retenu plusieurs champs dans un segment situé au nord de la



Quant aux problèmes des champs de petite taille et des pixels de chevauchement, on pourrait les résoudre en utilisant davantage de données au sol dans quelques petites surfaces, afin d'obtenir des estimations plus précises sur la superficie des récoltes dans une région. Vu l'étendue de la région où la pomme de terre est cultivée, on doit connaître la superficie des champs de pommes de terre à l'intérieur de dix à quinze segments, chaque segment ayant une superficie qui varie de cinq à huit kilomètres carrés. Toute la région serait alors classifiée le mieux possible, mais sans se servir de la classe subjective des pixels de chevauchement. La classification de chaque segment serait ensuite utilisée, avec les données au sol disponibles (les photographies aériennes prises en 1980, les données de l'EDA pour les années à venir), pour obtenir une relation de régression appliquée par la suite à l'estimation de toute la région afin de produire une estimation révisée (Hanuschak et al, 1979). On établirait aussi une estimation par le quotient, basée sur les estimations de toute la région obtenues par photographies aériennes et par données de satellite.

### 7.3 PRODUCTION DES ESTIMATIONS DE 1980 À PARTIR DES DONNÉES LANDSAT

On peut décrire l'estimation de la superficie plantée en pommes de terre, à partir des données satellite, comme un processus en trois étapes: l'obtention des segments de l'EDA, et, en dernier lieu, l'analyse des images.

Les données Landsat utilisées pour l'estimation étaient codées sur des bandes pour ordinateur et traitées sur le système de correction des images numériques par l'interpolation ( $\sin x/x$ ) pour la correction géométrique (Guertin et al, 1979) et par la correction radiométrique CAL 3 (Ahern et Murphy, 1978). Chaque bande pour ordinateur couvre quatre cartes du Système topographique national au 1:50,000<sup>e</sup> et comporte des pixels carrés de 50 m. Quatre bandes pour ordinateur sont nécessaires pour couvrir la région. Des données existantes avaient été commandées et elles devaient être livrées en mai 1980, alors que de nouvelles données avaient été commandées pour les passages du satellite au-dessus de la région entre la mi-juillet et la mi-août. La commande n'a posé aucun problème pour les données existantes, mais elle s'est avérée plus compliquée pour les données

terre. D'autres facteurs ayant contribué à cette marge d'erreur sont discutés ci-dessous.

## 7.2 AMÉLIORATION DES ESTIMATIONS

Bien que les travaux précédents aient connu un certain succès, on a identifié les sources potentielles d'erreur pour des applications nécessitant une précision supérieure à, disons, 85 %. Les principaux problèmes sont associés au choix subjectif des pixels de chevauchement des champs de pommes de terre, à la manipulation des champs de petite taille et à l'élimination des récoltes présentant des caractéristiques spectrales similaires à celles des pommes de terre.

En ce qui concerne ce dernier problème, l'idéal serait de connaître précisément la réflectance spectrale des pommes de terre et celle des autres récoltes tout au cours de la saison de croissance. À l'aide de ces renseignements, il serait alors possible d'identifier la fenêtre phénoménologique pendant laquelle les pommes de terre peuvent être distinguées avec certitude des autres récoltes. Malheureusement, un tel ensemble de données n'existe pas, bien que les connaissances actuelles sur les récoltes et les pratiques agricoles de la région fournissent des indications générales. Dans ce cas particulier, l'expérience des auteurs sur le terrain leur a permis de supposer que la date optimale pour distinguer les pommes de terre des autres récoltes dans cette région serait entre la mi-juillet et la mi-aout. Pour vérifier cette hypothèse et fournir une indication quant au degré de séparabilité des pommes de terre et des autres récoltes, on a analysé des données MSS (Landsat) sur bande d'ordinateur obtenues au dessus de la vallée de la rivière Saint-Jean le 8 aout 1975. La figure 1 indique les valeurs relatives de radiance dans chaque bande Landsat pour les pommes de terre, le maïs, les pois, le foin, le brocoli, le pâturage, le sarrasin, le sol nu et les céréales. On voit sur cette figure que les pommes de terre sont aisément séparables des autres récoltes, sauf les pois qui sont habituellement récoltés entre la mi-aout et la fin aout. Il semble donc que l'analyse des données recueillies vers la fin de la saison de croissance permettent la séparation et l'identification des pommes de terre.

pixels présentent un problème spécial, car, comme nous venons de le mentionner, ils chevauchent deux champs différents. La réflectance de ces pixels dépend de la proportion de la superficie du pixel couverte par chaque champ et de la réflectance du matériau superficiel dans chaque champ. Il est normalement très difficile de calculer le pourcentage de chaque type de matériau présent dans ces pixels. Toutefois, on a pu obtenir des estimations acceptables des superficies en modifiant la limite originale des zones de décision: on y a joint un second parallélogramme par apprentissage sur un certain nombre de ces pixels de chevauchement qui semblaient appartenir aux champs de pommes de terre. Pour choisir les pixels de chevauchement appropriés et les inclure dans la classe des pommes de terre, on s'est fondé sur une interprétation visuelle et subjective de la scène affichée à l'écran (les données de trois des quatre bandes spectrales étaient fusionnées pour former une image dont les couleurs ressemblaient à celle d'une émulsion de couleurs pour l'infrarouge).

L'estimation de la superficie pour toute la ceinture de la pomme de terre a nécessité moins de quatre heures d'utilisation du système CIA5. La localisation, l'affichage et l'analyse du secteur primaire et des sous-secteurs ont pris un peu plus d'une heure. La sélection, l'affichage et l'analyse des cinq sous-scènes choisies ultérieurement ont pris deux heures et demi, et une autre heure a été nécessaire pour localiser la frontière du Nouveau-Brunswick et éliminer les données appartenant à l'extérieur de la province.

Par rapport à la superficie totale plantée de pommes de terre qui avait été déterminée et mesurée à partir des photographies aériennes à basse altitude prises en même temps, les estimations obtenues par satellite en 1975 étaient précises à 95 % (c'est-à-dire qu'elles correspondaient à 95 % de la valeur véritable estimée) dans la sous-région contenant le site d'apprentissage, à 80 % dans la deuxième sous-région et à 88 % pour l'ensemble de la région d'étude. Lors d'essais répétés utilisant des champs d'apprentissage différents, la précision est passée de 85 à 97 % pour la région d'étude de primaire, alors que la précision pour l'ensemble de la province s'établissait à 84,5 %. Une partie de la marge d'erreur, dans l'estimation provinciale, était due au fait qu'un certain pourcentage des champs de pommes de terre se trouvaient à l'extérieur de la ceinture dite de la pomme de



graphies aériennes (échelle de 1" pour 832') de chaque segment échantillon-  
né. Les photographies provenaient de sources provinciales. La plupart  
d'entre elles avaient été prises en 1976. Ces agents recenseurs devaient  
visiter les agriculteurs exploitant des terres à l'intérieur de chaque seg-  
ment afin de leur présenter les photographies, pour qu'ils puissent identi-  
fier tous leurs champs de pommes de terre et de maïs<sup>3</sup>; les agents recen-  
seurs devaient noter les superficies indiquées par les agriculteurs. Le  
manuel des enquêteurs contenait une section sur les procédures à suivre,  
les procédures faisant partie de leur entraînement pour l'EDA.

## 7. ANALYSE DES DONNÉES OBTENUES PAR TÉLÉDETECTION

### 7.1 TRAVAUX PRÉCÉDENTS

Les travaux réalisés dans cette même région à l'aide de données de 1975 ont  
été rapportés ailleurs (Mosher et al) et la description détaillée de la mé-  
thodologie utilisée a été publiée (Ryerson et al, 1980). Dans les travaux  
de 1975, on a retenu une zone d'étude contenant environ 20 % de toutes les  
récoltes de pommes de terre de la province. À l'appui de cette étude, on a  
obtenu des données de terrain pour toute la région d'étude.

La région d'étude de 125 kilomètres carrés et deux sous-régions ont été lo-  
calisées sur l'écran couleurs du Système d'analyse d'images du CCT (CIAS -  
Goodenough, 1979). On a eu recours à une méthode très simple d'apprentis-  
sage dirigé pour dresser les données des pixels dans trois champs de pommes  
de terre sous forme de quatre histogrammes unidimensionnels. Un paralléli-  
pipède quadridimensionnel a été défini par les limites de chaque histogram-  
me et a servi à borner les zones de décisions. Tous les points qui se  
trouvent à l'intérieur du parallélipède ont été classés comme des pommes  
de terre, et ceux qui se trouvent à l'extérieur de cette région ont été  
classés sous la rubrique "autres".

Un des principaux problèmes rencontrés dans l'analyse des données était la  
classification adéquate des pixels situés sur la bordure des champs. Ces  
3 On devait aussi identifier les champs de maïs, car les travaux  
antérieurs en télédétection ont indiqué que le maïs pouvait être  
confondu avec les pommes de terre (Ryerson et al, 1980).

Dans le cadre de l'EDA au Nouveau-Brunswick, l'échantillon aréolaire a été jugé approprié pour obtenir les données au sol requises pour l'interprétation des données de télédétection. Cet échantillon a été composé en deux phases. Dans la première phase, les secteurs de dénombrement (SD) du recensement (lesquels comprenaient les sièges des fermes du recensement de 1976 - appelés SD agricoles du recensement) ont été stratifiés selon la superficie des champs de pommes de terre, l'importance du bétail et le nombre de porcs (données du recensement de 1976). À l'intérieur de chaque strate, on a choisi deux échantillons aléatoires simples par réplicat de secteurs de dénombrement. Chaque secteur échantillonné a été ensuite segmenté à l'aide de cartes, en superficie identifiable d'environ 5 à 8 kilomètres carrés, et on a choisi, comme échantillon aléatoire simple, un segment sur 10 par secteur de dénombrement. On a fourni aux agents recenseurs de l'EDA travaillant dans la région d'étude des agrandissements d'anciennes photo-

## 6. SAISIE DES DONNÉES ET ÉCHANTILLONNAGE AU SOL EN 1980

La région est densément boisée, de topographie variée et accidentée. Le drainage et la nature pierreuse du sol présentent certains problèmes dans cette région. On y compte quelque 70,000 hectares de terres agricoles amélorées, dont environ 20,000 servent habituellement à la culture de la pomme de terre. Les autres cultures importantes de ces régions sont les céréales, le foin et les légumes de traitement comme les pois, le brocoli et le chou de Bruxelles. La taille des terrains varie des parcelles de semences de 0.1 hectare aux champs de 40 hectares.

La région visée par l'étude est située dans la vallée supérieure de la rivière Saint-Jean, au Nouveau-Brunswick. Cette région prend naissance au sud de Woodstock, dans le comté Carleton, et suit la rivière Saint-Jean en direction nord-ouest sur environ 200 kilomètres, à travers le comté Victoria, jusqu'à Claire, dans le comté de Madawaska.

## 5. RÉGION D'ÉTUDE

satellite, on a retenu la date de publication de la seconde estimation.



L'enquête postale, l'enquête objective sur le rendement des pommes de terre (EORPT) et l'enquête descriptive sur l'agriculture (EDA).

Les questionnaires de l'enquête postale sont expédiés au début de juin à tous les agriculteurs répertoriés dans le registre des fermes dressé par la Division de la statistique agricole. Les réponses sont compilées par comté, et l'on obtient des estimations par comté en reliant la variation annuelle des superficies indiquées à la superficie totale de pommes de terre déterminée d'après le recensement pour chaque comté. Les estimations par comté sont ensuite additionnées afin de donner les estimations provinciales pour la fin juin.

L'enquête objective sur le rendement des pommes de terre est une enquête spécialisée, à la fois postale et descriptive, conçue pour estimer la superficie des récoltes de pommes de terre, ainsi que le rendement de la pomme de terre dans son habitat du Nouveau-Brunswick. L'enquête est réalisée à la mi-juillet sur un échantillon aléatoire de producteurs de pommes de terre obtenu du registre des fermes; les estimations des superficies de pommes de terre sont produites vers la mi-août.

L'enquête descriptive sur l'agriculture (EDA) est, comme son nom l'indique, une enquête descriptive et polyvalente conçue pour estimer les récoltes, le bétail et les dépenses agricoles au niveau provincial. L'EDA est une enquête à bases multiples réalisée à partir d'un échantillonage aléatoire d'agriculteurs choisis dans le registre des fermes et à partir d'un échantillon aléatoire et aléatoire de segments (parcelles de terre). Les agents recenseurs visitent les fermes échantillonnées à la fin juin et au début de juillet. Les estimations des superficies sont disponibles au début d'août. Chaque année, environ 20 % des segments sont renouvelés.

Au cours de la saison de croissance, on publie deux estimations de la superficie plantée en pommes de terre. La première, publiée au début de juin, est basée sur les résultats de l'enquête postale. La seconde estimation, publiée au début de septembre, est basée sur l'examen des estimations provenant des trois enquêtes et de discussions avec les autorités provinciales. Comme date cible pour la publication des estimations obtenues par

Au fil des ans, Statistique Canada a utilisé les données obtenues par les enquêtes postales annuelles comme intrants principaux pour son système d'estimation des récoltes. Bien que ces enquêtes soient relativement peu coûteuses et qu'elles peuvent être réalisées rapidement, elles sont limitées par les taux de réponses variables et par la non-représentativité toujours possible des répondants. Les enquêtes descriptives probabilistes ont été introduites au cours des années 1970, afin de pallier à certains de ces problèmes. Ces procédures font appel à des enquêteurs sur le terrain pour le dénombrement d'un échantillon aléatoire d'agriculteurs. En 1980, Statistique Canada a fondé ses estimations des superficies plantées en pommes de terre au Nouveau-Brunswick sur les résultats de trois enquêtes:

#### 4. SAISIE DES DONNÉES STATISTIQUES

Des données au sol précises sont requises pour deux raisons: en premier lieu, pour localiser les champs d'apprentissage étendus et, en second lieu, pour corriger toute erreur systématique dans la classification des données obtenues par satellite. Ces données au sol peuvent être obtenues par des recenseurs spécialisés travaillant sur le terrain, ou encore en utilisant une imagerie aérienne qui est interprétée par des analystes d'image.

Pour déterminer la superficie d'une récolte dans une région donnée, on compte le nombre de pixels se trouvant à l'intérieur de cette région et ayant été identifiés comme appartenant à la récolte. Un "apprentissage" additionnel peut être exécuté afin d'identifier les pixels "manqués" dans la classification initiale, ou encore pour réduire la confusion entre les récoltes, c'est-à-dire pour mieux séparer les récoltes dont la signature spectrale ressemble fortement à celle de la récolte d'intérêt.

Pour estimer la superficie des récoltes, on doit identifier chaque pixel qui appartient à un type de récolte donné. Des grands champs, que l'on sait appartenir au type de récolte qui nous intéresse sont localisés pour "l'apprentissage" du système. Les données de ces champs permettent d'identifier la signature particulière à cette récolte. Tous les pixels sont ensuite classés comme appartenant ou n'appartenant pas à cette récolte, selon leurs signatures spectrales.

### 3. TÉLÉDETECTION À L'AIDE DES SATELLITES LANDSAT

La télédétection consiste à mesurer les caractéristiques d'un objet à distance, habituellement à partir d'un aéronef ou d'un satellite. Avec les données captées par satellite, on peut obtenir rapidement la couverture complète de grandes superficies, et ce, à un coût relativement faible. Parmi les différents domaines d'application de ces données, mentionnons l'agriculture, l'exploitation forestière, l'utilisation des terres, la formation des glaces et la cartographie en général.

Les satellites américains de la série Landsat, lancés par la NASA, ont fourni les données pour cette expérience, et celles utilisées auparavant pour une expérience précédente au Nouveau-Brunswick. Chaque satellite Landsat fait le tour de la terre 14 fois par jour, en orbite héliosynchrone (ce qui permet ainsi de couvrir les différents points de la terre à la même heure solaire locale). La lumière réfléchie par le sol est enregistrée dans quatre bandes spectrales étroites, à l'aide du balayeur multispectral (MSS). Les données transmises au Canada sont captées à l'une des deux stations de réception situées, respectivement, à Prince Albert (Saskatchewan) et à Shoe Cove (Terre-Neuve). Chaque point de la terre est survolé tous les 18 jours par un satellite Landsat (si deux satellites sont utilisés, cette fréquence de couverture est de neuf jours).

Les données sont analysées par le Système d'analyse d'images du CCT (CIAS); ces données étant codées sur des produits standard comme les bandes d'ordonateur, chaque scène imagee couvrant une superficie de 25,600 kilomètres carrés. La plus petite unité de surface imagee est appelée élément d'image, ou pixel (de l'anglais "picture element"). Chaque pixel est identifié par une signature spectrale qui lui est propre, cette signature étant une mesure de la réflectance de la zone imagee dans les quatre bandes spectrales. La signature spectrale dépend des objets présents dans le pixel (rouges, cultures, etc.), chaque objet ayant une signature particulière. La signature des terres cultivées dépend de la structure des plantes, du type de sol apparent, de la maturité des récoltes, de la hauteur des plantes, de la densité des feuilles et de nombreux autres facteurs.



L'analyse des données obtenues par satellite a été faite en temps réel (c'est-à-dire presque instantanément) au CCT, car la plupart des travaux de préparation avaient été effectués avant l'acquisition des données. La zone d'échantillonnage retenue pour l'enquête descriptive sur l'agriculture (EDA) a fourni les données au sol nécessaires pour l'étalonnage du système; ces données ont permis d'établir des estimations par le quotient et par régression pour la correction des erreurs systématiques dans la classification par télédétection des champs de pommes de terre. Bien que cette démonstration n'ait pas eu lieu dans un cadre opérationnel, on aurait pu produire les estimations définitives moins de deux semaines après le passage du satellite au-dessus de la région d'étude.

La classification par satellite a été entravée par la présence de nuages (non-réponse au niveau du satellite) et par la similitude entre la signature spectrale des pommes de terre et celle d'autres objets présents dans la région étudiée. Le premier de ces problèmes s'est traduit par une perte de données, et sa correction a nécessité une certaine dose d'imputation. Quant au second problème, on l'a résolu en partie en corrigeant la classification et en utilisant des estimateurs par le quotient et par régression.

La comparaison des données recueillies au sol (par les enquêteurs), à partir des photographies aériennes des champs échantillonnés, a révélé que certains champs ont été omis par les enquêteurs de l'EDA, car les données sur ces champs n'étaient pas requises pour les besoins de l'EDA. Aussi a-t-on dû se servir des données tirées de photographies aériennes au lieu des données de l'EDA pour les estimations par satellite en 1980. En 1981, les procédures d'énumération de l'EDA ont été modifiées pour satisfaire tant aux exigences de l'EDA qu'à celles de la télédétection.

À la suite du succès rencontré par cette démonstration, l'expérience a été reprise en 1981. De plus, une expérience similaire a aussi été entreprise en 1981 afin d'estimer la superficie des récoltes de colza dans le district de Peace River, à la frontière de l'Alberta et de la Colombie-Britannique.

terre (Mosher et al, 1978; Ryerson et al, 1979; Ryerson et al, 1980). L'intérêt manifesté dans cette étude et dans d'autres travaux du CCT sur le projet de démonstration en temps réel utilisant les données recueillies par le satellite Landsat au cours de l'année-récolte 1980. Statistique Canada, qui est l'organisme fédéral chargé d'obtenir des données sur les récoltes, désirait comparer les estimations des superficies plantées obtenues par des méthodes traditionnelles avec les estimations obtenues par satellite pour une même région. On a retenu la pomme de terre comme récolte-cible, et la vallée de la rivière Saint-Jean comme région d'étude.

La télédétection par satellite présente plusieurs attraits pour Statistique Canada: précision accrue des estimations établies à partir d'enquêtes régulières (probablement à des niveaux plus locaux), allègement du fardeau imposé aux répondants (ce qui pourrait se traduire par une réduction du nombre de questionnaires et des questions) et possibilité de dresser des cartes pour de petites régions où poussent des récoltes spéciales, ceci afin de mieux planifier les méthodes d'échantillonnage.

La section suivante résume les principaux résultats de cette démonstration, et le reste de la communication indique en gros les méthodes de télédétection qui ont été utilisées dans ce projet. On décrit aussi le système actuel de saisie des données de Statistique Canada, la région visée par l'étude, l'échantillon au sol et la saisie des données sur le terrain, la vérification et l'analyse des résultats.

## 2. PRINCIPAUX RÉSULTATS

Les données recueillies par satellite ont servi à estimer la superficie des cultures de pommes de terre dans la vallée de la rivière Saint-Jean, au Nouveau-Brunswick. Ces estimations, extrapolées au niveau provincial, correspondaient à 2 % près aux estimations de 52,000 acres publiées par Statistique Canada. Ces estimations étaient basées sur les résultats de trois enquêtes indépendantes menées dans la province.



# UNE ESTIMATION PRÉCISE ET RAPIDE DE LA SUPERFICIE PLANTÉE EN POMMES DE TERRE GRÂCE À LANDSAT: RÉSULTATS D'UNE DÉMONSTRATION<sup>1</sup>

R.A. Ryerson, J.-L. Tamby, R.J. Brown

et

L.A. Murphy, B. McLaughlin<sup>2</sup>

Cette communication décrit les procédures et les résultats d'un projet conjoint de Statistique Canada et du Centre canadien de télédétection (CCT) visant à fournir une estimation rapide de la superficie plantée en pommes de terre au Nouveau-Brunswick, province qui est un important producteur de pommes de terre au Canada. Le projet a démontré que l'imagerie obtenue par satellite et combinée aux méthodes plus usuelles d'estimation des superficies plantées en pommes de terre peut alléger le fardeau des répondants, fournir rapidement des cartes de la distribution des cultures et permettre l'établissement d'estimations fiables pour les sous-régions.

## 1. INTRODUCTION

Les premières applications de la télédétection (par satellite) dans la vallée de la rivière Saint-Jean (Nouveau-Brunswick) par le Centre canadien de télédétection et le ministère de l'Agriculture du Nouveau-Brunswick ont démontré que les données obtenues par satellite pouvaient fournir des estimations précises et peu coûteuses de la superficie plantée en pommes de

<sup>1</sup> Communication initialement présentée au Quinzième symposium international sur la télédétection de l'environnement, Ann Arbor (Michigan), mai 1981.

<sup>2</sup> R.A. Ryerson et R.J. Brown, Centre canadien de télédétection (CCT), EMR; J.-L. Tamby, Division des méthodes d'enquête-entreprises (ce travail a été fait lorsque l'auteur était à la Division des méthodes d'enquête - institutions et agriculture), Statistique Canada; L.A. Murphy, Division de la statistique agricole, Statistique Canada; et B. McLaughlin, Division de la statistique agricole, Bureau régional de Turco (Nouvelle-Écosse), Statistique Canada.

## 6. RÉSUMÉ

Le premier recensement aura lieu en 1983, bien que ce n'est pas avant l'été 1984 que paraîtront les premières estimations de l'enquête remaniée. Pour cette raison, il sera nécessaire de mener à la fois l'ancienne enquête et la nouvelle enquête en 1983 si l'on veut disposer d'estimations pour cette année-là.

Bien qu'on ait conçu l'enquête de façon à ce qu'elle soit autopondérée, elle ne l'est pas complètement en raison des écarts temporels entre les recensements de 1981, la confection initiale des listes en 1983 et les recensements subséquents. Les formules d'estimation présentées à la section précédente tiennent compte de ces écarts temporels. Cependant, étant donné la grande simplicité de l'estimation autopondérée, nous avons proposé que l'on étudie son efficacité, une fois les données disponibles.

## BIBLIOGRAPHIE

- [1] Cochran, W.G. (1963), Sampling Techniques, John Wiley and Sons, New York.
- [2] Davidson, G. (1977), Redesign of the sour cherry, peach and grape objective yield surveys in the Niagara Peninsula, Survey Methodology 3, 38-61.
- [3] Jessen, R.J. (1955), Determining the fruit count on a tree by randomized branch sampling, Biometrics 11, 99-109.
- [4] Kish, L. (1965), Survey Sampling, John Wiley and Sons, New York.
- [5] Murthy, M.N. (1967), Sampling Theory and Methods, Statistical Publishing Society, Calcutta.
- [6] Singh, U. et Sukhatme, B.V. (1980), Sampling for estimating production of fruit crops, Sankhya C42, 17-30.

5.4 Estimations, par région, de la production fruitière totale et précision de ces estimations

On représente par  $\hat{X}_a^T$  le rendement réel (en tonnes de l'année précédente, dans la région a, et par  $\hat{Y}_a^T$  les estimations correspondantes de l'année courante. Puis, on estime  $\hat{Y}_a^T$  comme suit :

$$\hat{Y}_a^T = \hat{X}_a^T R_a \quad (5.4.1)$$

Le coefficient de variation est calculé de la façon suivante :

$$CV(\hat{Y}_a^T) = \frac{\hat{X}_a^T \{V(\hat{R}_a)\}^{\frac{1}{2}}}{\hat{X}_a^T R_a} \times 100\% = CV(\hat{R}_a) \quad (5.4.2)$$

5.5 Estimations de la production fruitière totale et précision de ces estimations

On représente par  $\hat{Y}^T$  la production fruitière totale dans les quatre régions, pour l'année courante, que l'on estime comme suit :

$$\hat{Y}^T = \sum_{a=1}^4 \hat{Y}_a^T \quad (5.5.1)$$

Le coefficient de variation est calculé de la façon suivante :

$$CV(\hat{Y}^T) = \frac{\{\sum_{a=1}^4 (\hat{X}_a^T)^2 V(\hat{R}_a)\}^{\frac{1}{2}}}{\hat{Y}^T} \times 100\% \quad (5.5.2)$$

$$V(\hat{y}_a) = (\hat{y}_{a1} - \hat{y}_{a2})^2 / 4 = D_{ya}^2 / 4 \quad (5.3.4)$$

$$V(\hat{x}_a) = (\hat{x}_{a1} - \hat{x}_{a2})^2 / 4 = D_{xa}^2 / 4 \quad (5.3.5)$$

$$\text{Cov}(\hat{y}_a, \hat{x}_a) = (\hat{y}_{a1} - \hat{y}_{a2})(\hat{x}_{a1} - \hat{x}_{a2}) / 4 = D_{ya} D_{xa} / 4 \quad (5.3.6)$$

où les indices inférieurs numériques correspondent au numéro de l'échantillon répété. Puis, on estime la variance,  $V(\hat{R}_a)$ , du taux de changement estimé,  $\hat{R}_a$ , de la façon suivante [1] :

$$V(\hat{R}_a) = \frac{1}{\hat{X}_a^2} \{ V(\hat{y}_a) - 2 \hat{R}_a \text{Cov}(\hat{y}_a, \hat{x}_a) + \hat{R}_a^2 V(\hat{x}_a) \}$$

$$= \left\{ \frac{S_{ya}}{D_{ya}} - \frac{S_{xa}}{S_{ya} D_{xa}} \right\}^2 S_{xa}^2$$

$$\text{ou } S_{ya} = \hat{y}_{a1} + \hat{y}_{a2}$$

$$D_{ya} = \hat{y}_{a1} - \hat{y}_{a2}, \text{ etc.}$$

$$(5.3.8)$$

On peut alors établir le coefficient de variation de  $\hat{R}_a$  comme suit :

$$CV(\hat{R}_a) = \frac{\hat{R}_a}{\{V(\hat{R}_a)\}^{\frac{1}{2}}} \times 100\% \quad (5.3.9)$$

## 5.3 Estimations, par région, du taux de changement et précision de ces estimations

On estime le taux de changement applicable à la production de la région a par rapport à l'année précédente, représenté par  $R_a$  comme suit:

$$R_a = \frac{\hat{Y}_a}{\hat{X}_a} \quad (5.3.1)$$

où  $\hat{Y}_a$  = le nombre total estimé de fruits commercialisables (pêches, griottes, tous les raisins ou raisins de la variété v) dans la région a, pour l'année en cours (ce nombre correspond à

$$\hat{Y}_a = \frac{1}{2} \sum_{i=1}^2 \hat{Y}_{ar} \quad (5.3.2)$$

dans le cas des pêches, des griottes et de tous les raisins de toutes les variétés, et à

$$\hat{Y}_a = \frac{1}{2} \sum_{i=1}^2 \hat{Y}_{arv} \quad (5.3.3)$$

dans le cas des raisins d'une variété donnée); et où  $\hat{X}_a$ ,  $\hat{X}_{ar}$ ,  $\hat{X}_{arv}$  représentent les estimations correspondantes de l'année précédente.

(Il est à noter qu'on peut omettre l'indice inférieur v, toutes les estimations étant calculées de la même façon, qu'il s'agisse des pêches, des griottes, de tous les raisins de toutes les variétés ou des raisins d'une variété donnée.)

On définit les variances de  $\hat{Y}_a$  et de  $\hat{X}_a$  ainsi que leur covariance comme suit:



$N_{arfb}^{83}$  = Le nombre total d'arbres ou de vignes (variété v), dans le verger ou le vignoble b, sur la ferme f, dans l'échantillon répété r, dans la région a, selon les cartes établies en 1982-1983;

$N_{arvf}^{83}$  = Le nombre total d'arbres ou de vignes (variété v), sur la ferme f, dans l'échantillon répété r, dans la région a, selon les listes établies en 1983;

$N_{ar}^{81}$  = Le nombre total d'arbres ou de vignes (toutes les variétés), sur la ferme f, dans l'échantillon répété r, dans la région a, selon le recensement de 1981 (accompagné de la liste-échantillon);

$N_a^{81}$  et  $N_a^{81}$  = Le nombre de tous les arbres ou de toutes les vignes dans la région a, selon le recensement de 1981 (voir le tableau-4).

TABLEAU 4: Nombre d'arbres ou de vignes ( $N_a^{81}$ ) par région, établi lors du recensement de 1981 et utilisé pour les enquêtes sur les fruits tendres

	REGION 1	REGION 2	REGION 3	REGION 4	TOTAL
PÊCHES	13,094	389,157	10,271	411,697	824,219
GRIOTTES	9,496	50,888	55,449	32,536	148,369
RAISINS	1,142,067	5,666,008	946,509	3,975,202	11,729,786

$n_{arfb}$

= le nombre d'arbres ou de vignes (variété v) qui ont fait l'objet d'un échantillonnage pendant l'année en cours, dans le verger ou le vignoble b, sur la ferme f, dans l'échantillon répété r, dans la région a (généralement,  $n_{arfb} = 4$  dans les enquêtes sur les griottes et les pêches et 5 dans l'enquête sur les raisins);

$n_{arvf}$

= le nombre de vergers ou de vignobles (où la variété v a fait l'objet d'un échantillonnage dans l'année en cours) sur la ferme f, dans l'échantillon répété r, dans la région a (généralement,  $n_{arvf} = 1$ , sauf dans le cas des unités choisies plus d'une fois dans le même échantillon répété, par exemple, les grosses fermes);

$n_{arv}$

= le nombre total de fermes distinctes (où la variété v a fait l'objet d'un échantillonnage dans l'année en cours), dans l'échantillon répété r, dans la région a;

$n_{arv}^*$

= le nombre total de vergers ou de vignobles (où la variété v a fait l'objet d'un échantillonnage dans l'année en cours), dans l'échantillon répété r, dans la région a

$n_{arvfb}^c$

= le nombre total, à ce jour, d'arbres producteurs ou de vignes productrices (variété v) dans le verger ou le vignoble b, sur la ferme f, dans l'échantillon répété r, dans la région a;

(c'est-à-dire,  $n_{arv}^* = \sum_{f=1}^F n_{arvf}$ );



## 5. FORMULES D'ESTIMATION

### 5.1 Estimations du nombre de fruits par arbre ou par vigne

On représente par  $y_t$  le nombre total de fruits commercialisables sur un arbre (vigne)  $t$ . Dans l'enquête sur les pêches, on estime  $y_t$  par  $\hat{y}_t$  nombre total de pêches commercialisables que l'on compte sur un arbre-échantillon  $t$ . Dans l'enquête sur les griottes, on estime  $y_t$  de la façon suivante :

$$\hat{y}_t = \hat{y}_{t\lambda} / p_\lambda$$

(5.1.1)

où  $\hat{y}_{t\lambda}$  est le nombre total de griottes commercialisables que l'on compte sur une ou des branches-échantillon  $\lambda$  de l'arbre-échantillon  $t$  choisi;

et  $p_\lambda$  représente les probabilités de sortie de la ou des branches-échantillon  $\lambda$ . Enfin, dans l'enquête sur les raisins, on estime  $y_t$  comme suit:

$$\hat{y}_t = \frac{N_t}{n_t} \sum_{\lambda=1}^n \hat{y}_{t\lambda}$$

(5.1.2)

où  $N_t$  est le nombre total de grappes de raisins sur la vigne-échantillon  $t$ ;

$n_t$  est le nombre de grappes de raisins qui ont été comptées sur la vigne-échantillon  $t$  (généralement,  $n_t = 5$ );

et  $\hat{y}_{t\lambda}$  est le nombre de raisins sur une grappe  $\lambda$  de la vigne-échantillon  $t$ .

#### 4. REMPLACEMENTS

Bien que l'agent recenseur doive s'efforcer de recenser les mêmes arbres, branches ou vignes les années subséquentes, cela peut s'avérer impossible dans certains cas (par exemple, lorsque des branches ont été coupées, des vignes ou des arbres arrachés ou détruits d'une autre façon). Si une branche de griot-tier a été coupée, on doit choisir une autre branche, sur le même arbre, en suivant la méthode décrite précédemment. Si cela ne peut se faire, on doit tout comme pour les pêchers et les vignes, choisir de façon aléatoire un nouvel arbre dans le même verger. Si tout le verger ou tout le vignoble a été détruit, on doit choisir un nouveau verger ou un nouveau vignoble, sur la même ferme, en suivant la méthode décrite à la section 2.4. Dans tous les cas, on fera le recensement des branches, arbres, vignes, vergers ou vignobles nouvellement sélectionnés; toutefois, les résultats obtenus ne serviront pas à l'établissement des estimations de l'année courante, mais bien seulement de l'année suivante, car seules les données apparues sont considérées.

Dans les cas (peu fréquents, espérons-le) où le producteur a cessé complètement la culture du fruit visé ou refuse de collaborer dès le contact initial, on doit choisir, pour chacune des enquêtes, un troisième échantillon répété de taille beaucoup moins importante, sans remplacement. Les méthodes à suivre pour choisir le verger ou le vignoble et les arbres, branches ou vignes-échantillon pour chaque remplacement sont les mêmes que celles qui sont décrites dans les paragraphes précédents. Chaque année, on recense ces branches, ces arbres et ces vignes, mais les données obtenues ne servent à établir les estimations que lorsqu'il est nécessaire d'inclure une de ces unités dans l'échantillon. L'effectif de l'échantillon de remplacement, par région, figure au tableau 3.

Tableau 3 : Effectif de l'échantillon de remplacement, par région, pour les enquêtes sur les fruits tendres

	RÉGION 1	RÉGION 2	RÉGION 3	RÉGION 4	TOTAL
PÊCHES	1	2	1	3	7
GRIOTTES	1	2	2	2	7
RAISINS	1	3	1	2	7



## 2.6.2 Enquête sur les raisins

Comme pour les griottes, il est également impossible de compter tous les raisins commercialisables sur une vigne-échantillon. Par conséquent, pour estimer le total des raisins, on compte les grappes de raisins (celles qui ont plus de 5 raisins) et on choisit au hasard, sans remplacement, 5 grappes dont on compte les fruits un à un. Comme pour les autres enquêtes, les vignes sont marquées pour identification future, les mêmes unités étant recensées d'une année à l'autre.

## 3. COLLECTE DES DONNÉES

Le recensement lui-même a lieu environ quatre semaines avant la récolte, chaque année. Il est très important que les vignes, les branches et les arbres qui ont été sélectionnés, de même que les vergers et les vignobles, soient bien identifiés afin de permettre aux agents recenseurs de faire leur travail dans le court laps de temps dont ils disposent. Les agents recenseurs doivent compter tous les fruits commercialisables (c'est-à-dire, tous les fruits à l'exception des fruits de qualité inférieure non parvenus à maturité ou des fruits endommagés qui ne seront pas récoltés) sur les pêcheurs-échantillon et sur les branches de griottiers qui ont été sélectionnées. On compte tous les fruits d'un pêcheur principalement parce que la distribution des fruits sur cet arbre tend à être beaucoup plus inégale que celle des fruits sur un griottier [2], ce qui exclut la possibilité de simplement recenser des branches-échantillon.

Pour l'enquête sur les raisins, on compte tous les raisins composant les cinq grappes choisies, à l'exception des fruits de qualité inférieure. Comme les raisins sont souvent groupés de façon très serrée sur la grappe, il faut, la plupart du temps, récolter les fruits. Cette raison, ajoutée à des contraintes temporelles, fait qu'il est impossible de recenser toute la vigne-échantillon.

identification future, car les mêmes unités sont recensées d'une année à l'autre. (Les années subséquentes, si une vigne ou un arbre échantillonné a été détruit ou arraché ou qu'il est mort, on choisit un arbre ou une vigne de remplacement qu'on inclut dans le recensement. Cependant, les données sur cet arbre ou cette vigne de remplacement ne servent à établir les estimations qu'à partir de la deuxième année.) En outre, chaque année, on recompte les vignes ou les arbres producteurs faisant partie des vignobles ou des vergers sélectionnés afin de rendre compte de l'évolution de l'industrie. (Il est à noter que, dans le cas de l'enquête sur les raisins, on compte seulement les vignes de la variété échantillonnée, dans chaque vignoble.)

## 2.6 Plan d'échantillonnage - 4<sup>e</sup> degré

Cette étape ne s'applique qu'aux enquêtes sur les griottes et les raisins.

### 2.6.1 Enquête sur les griottes

Il est techniquement impossible de compter toutes les griottes sur un arbre sélectionné. Afin d'estimer le nombre total de fruits destinés à la vente, on choisit une ou plusieurs branches-échantillon, avec probabilité proportionnelle à la surface de la coupe transversale d'une branche. Jessen [3] décrit cette façon de choisir une branche. Cela consiste à sélectionner une branche au point initial (ou primaire) de ramification du tronc, avec probabilité proportionnelle à la surface de la coupe transversale, et à suivre la branche choisie jusqu'au prochain point de ramification. On répète cette opération jusqu'à ce que la surface de la coupe transversale d'une branche choisie subégalement soit égale à au moins cinq pour cent et à au plus quinze pour cent de la surface totale cumulative des coupes transversales des branches primaires. Comme il n'est pas toujours possible de choisir une telle branche, il faut, dans certains cas, recenser deux branches. La branche choisie dans chaque arbre-échantillon est marquée pour identification future, étant donné que la même unité est recensée d'une année à l'autre.

Les pêches et les griottes, l'agent recenseur demande au producteur de l'aider à énumérer tous ses vergers et à établir la taille (c'est-à-dire, le nombre d'arbres) de chacun. Dans le cas de l'enquête sur les raisins, il lui demande de dresser une liste similaire pour chacune des trois variétés de raisins qui nous intéressent, nommément Concord, DeChauzac et "Autres". (Il est à noter que, certaines variétés étant cultivées ensemble, un vignoble peut figurer sur plusieurs listes. Cependant, la taille déclarée du vignoble pour une variété particulière est fonction du nombre de vignes de cette variété seulement.)

Dans le cas des enquêtes sur les pêches et les griottes, on choisit, dans chaque ferme-échantillon, un verger avec probabilité proportionnelle à la taille. Dans le cas de l'enquête sur les raisins, on choisit de façon indépendante un vignoble pour chacune des trois variétés effectivement cultivées sur la ferme, également avec probabilité proportionnelle à la taille. Souignons à quel point il est important de suivre ces méthodes avec soin afin de ne pas compromettre la validité des estimations. Il faut surveiller de près l'étape de la sélection afin de s'assurer qu'aucun biais n'est introduit en faveur des petits vergers ou des vignobles à variété unique, bien entendu, plus faciles à recenser.

Afin d'éviter un chevauchement des vergers ou des vignobles faisant partie des fermes tirées dans les deux échantillons répétés ou tirées plus d'une fois dans le même échantillon répété, on choisit en même temps tous les vergers ou les vignobles pour une ferme donnée, en se servant d'une méthode d'échantillonnage systématique, avec probabilité proportionnelle à la taille. Suite à cette sélection, on procède, de façon aléatoire, à l'inclusion des vergers ou des vignobles dans les échantillons répétés. (Il est à noter que, dans l'enquête sur les raisins, pour les fermes figurant dans les deux échantillons répétés, deux vignobles de chaque variété cultivée.)

## 2.5 Plan d'échantillonnage - 3<sup>e</sup> degré

Une fois la sélection d'un verger ou d'un vignoble terminée, on procède au comptage des vignes ou des arbres producteurs, et on tire un échantillon aléatoire simple, sans remplacement, de quatre arbres producteurs ou de cinq vignes productrices. Les vignes ou les arbres choisis sont marqués pour

## 2.3 Plan d'échantillonnage - 1er degré

Pour chaque enquête, nous avons tiré systématiquement, dans chaque région, deux échantillons répétés et indépendants de fermes (afin d'obtenir un échantillon représentatif), avec probabilité proportionnelle au nombre total d'arbres ou de vignes déclaré par l'exploitant dans le recensement de 1981. Les effectifs totaux des deux échantillons répétés, par région, figurent dans le tableau 1. Étant donné que les deux échantillons répétés sont tirés de façon indépendante et que les fermes plus importantes ont plus de chance d'être choisies, on s'attend à un certain chevauchement entre les échantillons répétés. En fait, certaines fermes sont tellement grosses qu'elles ont non seulement la garantie de faire partie de l'échantillon mais également la possibilité de figurer plus d'une fois dans le même échantillon répété [4]. Chaque inclusion dans l'échantillon est considérée comme un événement distinct, et chaque verger ou vignoble est tiré sans remplacement chaque fois qu'une ferme est choisie. Par conséquent, le nombre réel de fermes distinctes faisant partie de l'échantillon se trouve diminué, comme on peut le voir au tableau 2.

**TABLÉAU 2:** Nombre total de fermes distinctes faisant partie de l'échantillon, par région

	RÉGION 1	RÉGION 2	RÉGION 3	RÉGION 4	TOTAL
PÊCHES	3	22	3	27	55
GRIOITES	4	17	15	10	46
RAISINS	4	30	4	20	58

## 2.4 Plan d'échantillonnage - 2<sup>e</sup> degré

À partir du deuxième degré, une certaine partie du travail d'échantillonnage se fait sur le terrain. Une fois que le contact initial avec le fruiticulteur est établi (au printemps de 1983), il est important de consentir tous les efforts nécessaires pour obtenir sa coopération. Dans le cas des enquêtes sur



Région 3: Villies de Pelham et de Thorold dans la municipalité régionale de Niagara (Townships 11 et 12, dans le comté 29)

Région 4: Villies de Niagara Falls et de Niagara-on-the-Lake, dans la municipalité régionale de Niagara (Townships 3 et 10, dans le comté 29)

Étant donné la demande croissante d'estimations de la production fruitière par aire géographique, nous avons tiré un échantillon distinct de fermes pour chacune des quatre régions et cherché à répartir les effectifs de l'échantillon (c'est-à-dire, le nombre de fermes choisies) de façon optimale entre les régions. Cependant, étant donné l'effectif exceptionnellement réduit de la population dans certaines régions (voir le tableau 1), nous avons fait un compromis entre une répartition proportionnelle, une répartition optimale et une règle établissant un "minimum de 2 fermes par région, par échantillon répété". Cette dernière règle nous a semblé appropriée, puisqu'elle réduit la possibilité d'obtenir une non-réponse totale dans un échantillon répété donné (comme cela pourrait être le cas si on ne choisissait qu'une seule ferme par échantillon répété). Le nombre d'arbres ou de vignes, dans chaque ferme, a servi de variable pour mesurer la taille de la ferme, aux fins de la répartition de l'échantillon et de la sélection de celui-ci, avec probabilité proportionnelle à la taille, aux 1<sup>er</sup> et 2<sup>e</sup> degrés. Des résultats antérieurs [6] indiquent que d'autres variables de substitution (telles que la superficie cultivée) ne sont probablement pas meilleures que la variable fondée sur le nombre d'arbres pour mesurer la taille de la ferme.

TABLEAU 1 : Effectifs de la population et de l'échantillon, par région, pour les enquêtes sur les fruits tendres

RÉGION 1	ÉCHAN- TILLON	POPU- LATION	RÉGION 2	ÉCHAN- TILLON	POPU- LATION	RÉGION 3	ÉCHAN- TILLON	POPU- LATION	RÉGION 4	ÉCHAN- TILLON	POPU- LATION	TOTAL
15	4	4	198	22	15	4	195	40	30	423	552	60
67	4	4	275	32	46	4	164	12	22	552	56	62



## 2.1 Population cible, bases de sondage et effectif total de l'échantillon

La population cible des trois enquêtes sur le rendement prévu comprend tous les fruiticulteurs commerciaux de la péninsule de Niagara. Le MAAO a défini les fruiticulteurs commerciaux comme étant ceux qui ont déclaré plus de 200 - pêcheurs, 200 griottiers ou 5,000 vignes lors du recensement des fruiticulteurs (arbres fruitiers et vignes) de 1981. À partir de cette définition, nous avons établi une base de sondage distincte pour chacune des trois enquêtes. Les listes servant aux enquêtes sur le rendement en pêches, en griottes et en raisins comptent respectivement 423, 145 et 552 fruiticulteurs commerciaux. Les contraintes budgétaires du MAAO ont servi à déterminer l'effectif total de l'échantillon (nombre de vergers et de vignobles à recenser) pour chaque enquête, soit environ 60 vergers de pêches, 55 vergers de griottiers et 155 vignobles. Dans le cas de l'enquête sur les raisins, comme toutes les variétés de raisins qui nous intéressent doivent être échantillonnées sur une ferme choisie, il est impossible de connaître d'avance l'effectif final de l'échantillon. Cependant, en nous fondant sur le recensement des vignobles de 1981, nous estimons que 62 fermes fourniront un échantillon d'environ 155 - vignobles.

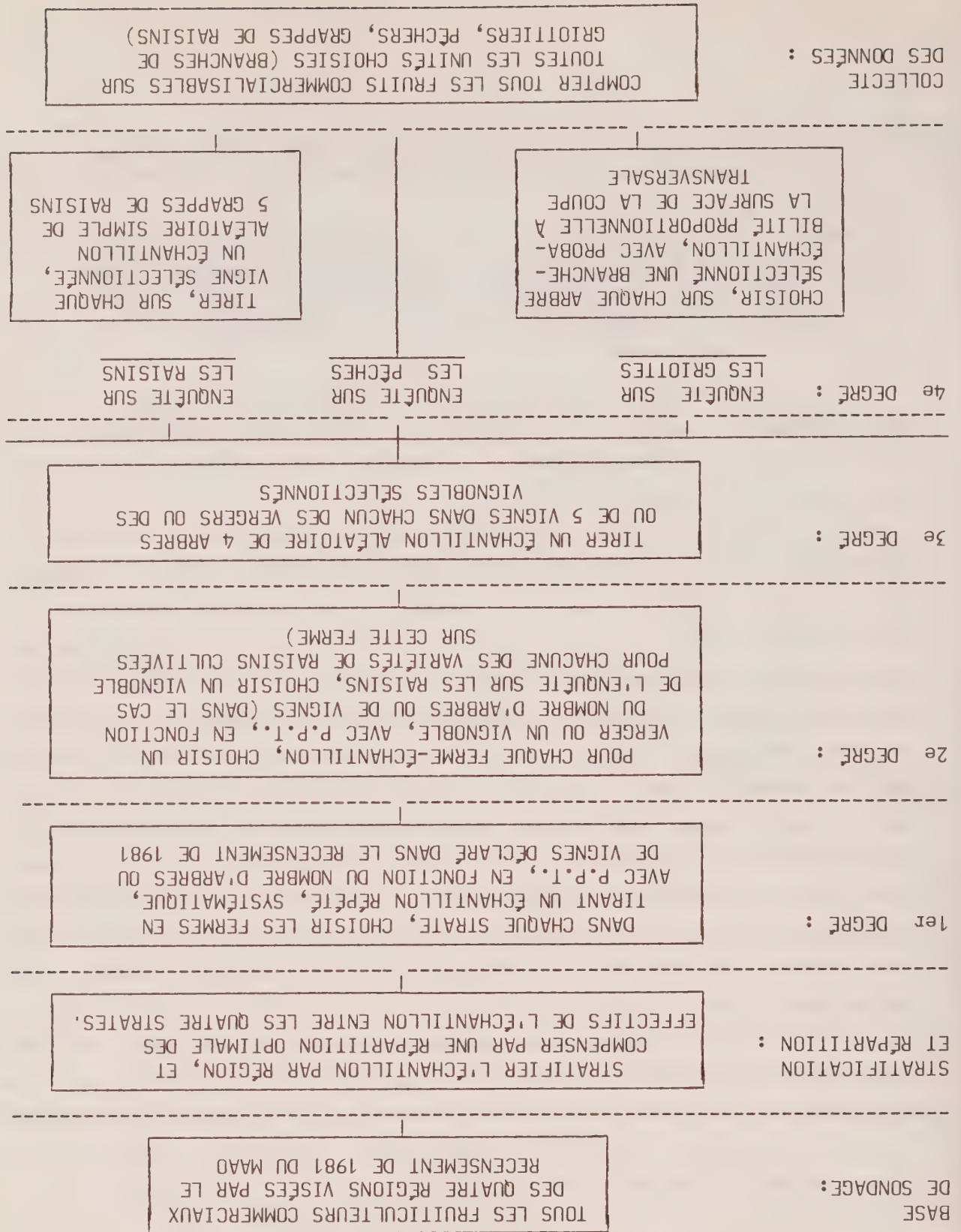
## 2.2 Stratification et répartition des effectifs de l'échantillon entre les régions

La péninsule de Niagara a été divisée en quatre régions, pour lesquelles il faut faire des estimations distinctes. Les quatre régions ont été définies comme suit (selon les limites établies lors du recensement de 1976) :

Région 1: Ville de Grimsby dans la municipalité régionale de Niagara et township de Saltfleet dans la municipalité régionale de Hamilton-Wentworth (Township 8, dans le comté 29, et township 4 dans le comté 17)

Région 2: Villes de St. Catharines et de Lincoln dans la municipalité régionale de Niagara (Townships 5 et 9, dans le comté 29)

Figure 1. Plan d'échantillonnage des enquêtes sur le rendement prévu en fruits tendres



étaient complètes et précises et qu'elles contenaient suffisamment de renseignements pour constituer la base de sondage des enquêtes sur les fruits tendres. Par conséquent, nous utiliserons dorénavant, pour les enquêtes sur le rendement en pêches, en griottes et en raisins, trois échantillons indépendants tirés des listes du recensement mené par le MAAO en 1981.

Les trois enquêtes ont pour objet de prévoir la quantité totale de fruits effectivement vendue (pour la vente fraîche ou la transformation). Nous établissons ces prévisions en estimant le rapport entre le nombre de fruits destinés à la vente pour l'année courante et le total correspondant pour l'année précédente et appliquons le ratio obtenu à la quantité de fruits, en tonnes, effectivement vendue l'année précédente (donnée fournie par le comité de la statistique sur les fruits et les légumes de l'Ontario). Ainsi, nous supposons qu'il y a une corrélation importante entre le poids des fruits et le nombre de fruits. Étant donné l'écart temporel entre les enquêtes et la récolte, nous supposons en outre que toute perte de fruits entre ces deux événements est constante d'une année à l'autre.

## 2. VUE GÉNÉRALE DU PLAN D'ÉCHANTILLONNAGE

Les échantillons des trois enquêtes sur le rendement prévu ont été tirés de façon indépendante, conformément à un plan d'échantillonnage stratifié (selon les aires géographiques), presque complètement autopondéré (tous les arbres et toutes les vignes ont à peu près la même chance d'être choisis), à plusieurs degrés, avec échantillons répétés et avec p.p.t. (les fermes de même que les vergers et les vignobles sont choisis selon une probabilité proportionnelle à leur taille). Un schéma du plan d'échantillonnage est présenté à la figure 1. Il est à noter que, comme les variables de pondération sont recueillies à des moments différents, le plan d'échantillonnage n'est pas parfaitement autopondéré.

## REMANIEMENT DE L'ENQUÊTE SUR LE RENDEMENT PRÉVU EN FRUITS TENDRES

## DANS LA PÉNINSULE DE NIAGARA

J. Kovar<sup>1</sup>

Des enquêtes sur le rendement prévu en pêches, en griottes (cerises acides) et en raisins sont menées chaque année, depuis 1964, dans la péninsule de Niagara; elles servent à prévoir l'amplitude des variations, d'une année à l'autre, dans la production de fruits destinés à la vente. Il est essentiel d'établir à temps les estimations afin de permettre à l'Office de commercialisation des fruits tendres de l'Ontario et à l'Office de commercialisation des raisins de l'Ontario de dresser leurs plans de commercialisation bien avant la récolte. Dans le présent rapport, nous expliquons les principales modifications apportées à l'enquête par suite du deuxième remaniement effectué en 1982. Nous examinons plus particulièrement le plan d'échantillonnage, la collecte des données et les formules d'estimation.

## 1. INTRODUCTION

Lors du premier remaniement, en 1974, il avait été décidé de changer la base de sondage de l'enquête, soit de se fonder sur des aires géographiques plutôt que sur une liste, principalement parce qu'on ne disposait d'aucune liste complète des fruiticulteurs commerciaux dans la péninsule de Niagara. Toutefois, en 1981, le ministère de l'Agriculture et de l'Alimentation de l'Ontario (MAAO) a fait le recensement des fruiticulteurs (arbres fruitiers et vignes). Avec les données ainsi obtenues, nous pouvons remanier l'enquête pour la deuxième fois afin de tenir compte des changements survenus dans l'industrie au cours des huit dernières années. Suite à des entretiens avec le MAAO, il a été décidé que les listes du recensement des fruiticulteurs

<sup>1</sup> J. Kovar, Division des méthodes d'enquêtes-entreprises, Statistique Canada. Ce travail a été fait lorsque l'auteur était à la Division des méthodes d'enquête-institutions et agriculture, Statistique Canada.

Table 1

a)		Mois occupé						Mois où les données sont présentes						b)		Mois en Chômage						Mois où les données sont présentes													
		0	1	2	3	4	5	6			0	1	2	3	4	5	6			0	1	2	3	4	5	6			0	1	2	3	4	5	6
1	52.23	3.22	1.93	2.21	2.45	3.55	34.41		51.48	3.04	2.17	3.36	4.22	35.73		49.26	4.15	3.16	4.83	38.60		46.31	6.18	5.62	41.89		51.40	8.32	40.28		52.87	47.13			
2	91.33	6.01	2.66						89.00	6.42	2.81	1.77				89.28	5.76	2.17	1.43	1.36		91.23	4.46	2.08	1.13	0.72	0.38		91.43	8.57					
3	89.00	6.42	2.81	1.77																															
4	89.28	5.76	2.17	1.43	1.36																														
5	91.23	4.46	2.08	1.13	0.72	0.38																													
6	92.43	3.63	1.65	0.93	0.57	0.43	0.35																												



- [7] Rubin, D.B. Inference and Missing Data. Biometrika, 1976, 63, 581-592.
- [8] Smith, R.E. and Vanski, J.E. Gross Change Data: The Neglected Data Base. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 132-150.
- [9] Stasny, E.A. Estimating Gross Flows in Labor Force Participation Using Information From Individuals With Incomplete Classifications. Technical Report 272, Department of Statistics, Carnegie-Mellon University, Jan. 1983.
- [10] Statistique Canada, Guide d'utilisation des données de l'enquête sur la population active. No 71-528 au catalogue, hors série, juillet 1979.
- [11] Statistique Canada, Méthodologie de l'enquête sur la population active du Canada 1976. No 71-526 au catalogue. Hors série. Octobre 1977.
- [12] Wong, F. A Technique To Correct The Response Rate In the 4x4 Labour Force Gross Flow Matrix. Technical Report, Statistics Canada, 1983.

Le but d'estimer des flux bruts entre les catégories de la population active serait certainement différente de l'enquête sur la population active. Ainsi, les données longitudinales de l'enquête ne sont pas idéales pour les estimations des flux bruts. Les données, cependant, sont disponibles, et si elles peuvent être utilisées pour fournir des estimations raisonnables des flux bruts, alors on produit des renseignements supplémentaires utiles à un coût relativement peu élevé.

## RÉFÉRENCES

- [1] Bishop, Y.M.M., Fienberg, S.E. et Holland, P.W. Discrete Multivariate Analysis: Theory and Practice. The MIT Press, 1975.
- [2] Deming, W.E. et Stephan, F.F. On A Least Squares Adjustment Of a Sampled-Frequency Table When The Expected Marginal Totals Are Known. Annals of Mathematical Statistics, 1940, 11, 427-444.
- [3] Fienberg, S.E. et Tanur, J.M. The Design et Analysis of Longitudinal Surveys: Controversies and Issues of Cost And Continuity. Technical Report 289, Department of Statistics, Carnegie-Mellon University, May 1983.
- [4] Kalachek, E. Longitudinal Surveys and Labor Market Analysis. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office, 1979, 160-189.
- [5] Macredie, I.D. and Vevers, R. Discussion on Smith, R.E. and Vanski, J.E. Gross Change Data: The Neglected Data Base. Data Collection, Processing and Presentation: National and Local (Counting the Labor Force, Appendix, Volume II), U.S. Government Printing Office 1979, 153-158.
- [6] Paul, E.C. et Lawes, M. Caractéristiques des ménages répondants et non-répondants dans l'enquête sur la population active du Canada. Techniques d'enquête, 1982, 8, 53-93.

Il est clair que le problème d'obtenir de bonnes estimations des flux bruts à partir des données de l'enquête sur la population active n'est pas un problème simple. L'enquête est organisée pour fournir des données pour la production des estimations mensuelles de l'activité sur le marché du travail, et non des estimations des flux bruts. Une enquête organisée spécifiquement dans

Lorsqu'on a appliqué ces modèles aux données de l'enquête sur la population active provenant d'un seul panel, Stasny (1983) a trouvé que le modèle où la probabilité de perdre le statut d'activité d'un mois dépend du statut d'activité correspond raisonnablement aux données pour toutes les matrices de flux bruts à l'exception de la matrice des mois 1-2. Pour les données du mois 1 au mois 2, la probabilité de la perte du statut d'activité d'un mois semble dépendre du mois. Ceci peut être dû au fait qu'il y a un taux de non-réponse plus élevé le premier mois qu'un panel est inséré dans l'enquête. Nous croyons qu'il vaudrait la peine d'appliquer ce genre de modèle à des données supplémentaires de l'enquête sur la population active pour vérifier si l'on obtient des résultats semblables avec d'autres panels.

n'utilise que les renseignements des individus qui ont répondu les deux mois. Il y a également des renseignements disponibles des individus qui n'ont répondu qu'un seul de ces deux mois. Stasny (1983) présente une méthode d'estimation des flux bruts d'un mois à l'autre qui emploie les renseignements disponibles des individus qui ont répondu seulement un de ces deux mois et qui peut être utilisée lorsque la non-réponse est reliée au temps ou au statut d'activité. Dans cette méthode, nous considérons les données de flux bruts observées comme le résultat final d'un processus à deux étapes. Dans la première étape du processus, que nous n'avons pas l'occasion d'observer, les individus sont distribués entre les 16 cellules de la matrice des flux bruts conformément à un schéma d'échantillonnage multinomial simple. Ensuite, dans la deuxième étape, chaque individu possède une probabilité quelconque de perdre son statut d'activité soit dans le mois  $t-1$  ou dans le mois  $t$ . On peut établir le modèle de la probabilité de perdre le statut d'activité d'un mois selon le mois, ou le statut d'activité, ou les deux. Les estimations de probabilité maximale pour les paramètres de la distribution multinomiale de la première étape et les probabilités de perdre le statut d'un mois sont obtenues au moyen de méthodes itératives.

Considérons les probabilités sous-jacentes aux pourcentages observés présentés dans la partie a) du tableau 1. Soit

$\pi_i$  = la probabilité qu'un individu soit occupé pendant  $i$  de 6 mois pour  $i = 0, 1, \dots, 6$ .

En postulant que la non-réponse se produit au hasard, les probabilités correspondantes à la première colonne de ce tableau peuvent s'écrire :

$P$  (qu'on observe 0 mois occupé sur 6-k mois de réponse)

$$= \sum_{j=1}^6 \pi_j / \binom{j}{k}, \text{ pour } k = 0, 1, \dots, 5. \quad (13)$$

A noter que ces probabilités augmentent de la première à la dernière ligne de la colonne.

De même, on peut montrer que, si les données manquent au hasard, alors les probabilités sous-jacentes doivent augmenter de façon descendante dans chaque colonne dans les deux tableaux. La première colonne de chaque tableau dévie de ce comportement de façon très nette. Dans les deux cas, les pourcentages observés diminuent tout au long des quatre premières lignes du tableau et augmentent dans les deux dernières. Il ne semble pas probable que les variations d'échantillonnage par elles-mêmes puissent être responsables d'un tel comportement dans les deux tableaux. Par conséquent, il semble bien que la non-réponse ne se produise pas au hasard.

Bien entendu, cette analyse n'est basée que sur un seul panel des données de l'enquête sur la population active. Cependant, dans une étude plus considérable utilisant des données de 1980 et 1981, Paul et Lawes (1982) ont également trouvé des preuves d'une relation entre le statut d'activité et la non-réponse. Par conséquent, il est nécessaire d'envisager des méthodes d'estimation des flux bruts qui ne nécessitent pas le postulat que la non-réponse se produit au hasard.

La méthode que propose Statistique Canada pour l'estimation des flux bruts



et à l'extérieur de celle-ci. Une autre possibilité qui devrait être prise en considération est l'élimination des poids mensuels aux fins de l'estimation des flux bruts, et le calcul d'un poids longitudinal pour chaque individu inclus dans l'échantillon de l'enquête sur la population active dans l'un ou l'autre des deux mois.

En tant que statisticiens, nous acceptons facilement des estimations des flux bruts dont les totaux marginaux ne correspondent pas aux totaux de l'activité sur le marché du travail publiés mensuellement; cependant, nous sommes conscients des problèmes qui pourraient être soulevés si on publiait des estimations des flux bruts non convergentes avec les totaux mensuels. Néanmoins, on ne devrait pas postuler automatiquement que les estimations mensuelles sont justes et que le problème se situe uniquement dans les estimations des flux bruts. Comme nous l'avons noté dans la section 3.5, la matrice des flux bruts est corrigée pour tenir compte des erreurs de classification. Les estimations mensuelles, cependant, ne sont pas corrigées pour le biais d'erreur de classification. En conséquence, lorsque l'ajustement proportionnel itératif est utilisé pour corriger la matrice des flux bruts pour la faire correspondre aux totaux mensuels, on change la matrice pour la faire converger avec des valeurs biaisées. Nous croyons qu'il serait plus approprié de faire face au problème des erreurs de classification du statut d'activité dans les données mensuelles là où elles se produisent plutôt qu'uniquement dans les estimations des flux bruts.

## 5. NON-RÉPONSE ET ESTIMATION DES FLUX BRUTS

La méthode proposée par Statistique Canada d'estimation des flux bruts corrigée la non-réponse en ajustant les poids des répondants en fonction de l'échantillon. Cette méthode de traitement de la non-réponse est correcte si les données manquantes sont distribuées au hasard (voir Rubin, 1976). Pour explorer le postulat de la répartition au hasard de la non-réponse, nous avons utilisé un fichier longitudinal pour un seul panel pour produire les données du tableau 1. Ce tableau indique les pourcentages non pondérés des individus interviewés occupés ou en chômage entre 0 à 6 mois selon le nombre de mois où ils ont répondu à l'enquête.



active n'est pas conçue pour l'estimation de nombres de personnes à l'extérieur de la population étudiée. Si nous voulons obtenir des estimations raisonnables pour les cellules d'entrée et de sortie de la matrice, il peut devenir nécessaire d'inclure des individus de l'extérieur de la population étudiée dans l'échantillon de l'enquête sur la population active ou d'utiliser un échantillon spécial supplémentaire.

Dans la section 4.3, nous avons vu que les surestimations des cellules d'entrée et de sortie pouvaient résulter de mouvements des individus entre une strate dont la population a augmenté et une autre dont la population a diminué. Le fait que ce sont les mouvements entre les strates qui ont causé les problèmes résulte des postulats simplificateurs qui ont été posés. Nous avons postulé que l'échantillon final a été choisi au hasard à l'intérieur de chaque strate. Par conséquent, les poids attribués aux individus inclus dans l'échantillon provenant d'une même strate étaient égaux. Si, à la place, nous avions supposé que les strates avaient été divisées en grappes et que les échantillons aléatoires d'individus avaient été choisis dans ces grappes, alors tous les individus inclus dans l'échantillon provenant d'une même grappe se seraient vu attribuer le même poids et la surestimation aurait été produite par les mouvements entre les grappes.

Pour corriger cette surestimation et les sous-estimations correspondantes, directement lorsque les échantillons finals sont choisis au hasard de l'intérieur des strates, nous aurions besoin d'estimations du nombre de sujets qui se sont déplacés entre chaque paire de strates lorsque la population d'une strate a augmenté et que celle d'une autre a diminué. Si les échantillons finals sont choisis au hasard à l'intérieur de grappes, des estimations semblables seraient nécessaires pour chaque paire de grappes. Ceci exige une quantité considérable de renseignements. Une complication supplémentaire est que, en pratique, les corrections proportionnelles appliquées aux poids rendent possible l'attribution de poids différents aux divers membres d'un même ménage.

Comme il a été suggéré plus tôt, si des individus à l'extérieur de la population étudiée étaient inclus dans l'échantillon, nous pourrions obtenir directement des estimations des mouvements à l'intérieur de la population étudiée

Enfin, nous notons que dans la matrice des flux bruts 2x2 présentée ci-haut la cellule Dans-la-population à Dans-la-population doit contenir une sous-estimation égale à la surestimation de la cellule de sortie. Quelle que soit la taille de cette surestimation, elle est étalée sur les neuf cellules Dans-la-population à Dans-la-population de la matrice des flux bruts 4x4. De plus, la taille de la surestimation est petite en comparaison avec la taille totale des neuf cellules Dans-la-population.

#### 4.4 Commentaires sur la méthode d'estimation des flux bruts proposée

Les résultats décrits dans les deux sous-sections précédentes illustrent des problèmes de la méthode proposée pour le traitement des différences des poids entre les mois aux fins de l'estimation des flux bruts. Ces résultats ne constituent pas une surprise pour les chercheurs de Statistique Canada. Grâce à leur expérience des méthodes et des données de l'enquête sur la population active, ils savaient que les mouvements des individus à l'intérieur de la population pourraient expliquer une partie de la surestimation des cellules d'entrée et de sortie de la matrice des flux bruts. Les résultats obtenus en basant le processus sur un modèle confirment cette opinion et précisent dans quelle mesure les mouvements des individus influencent les estimations. De plus, l'établissement du modèle a fait surgir un problème dont Statistique Canada n'était pas conscient: la sous-estimation compensatoire étalée à travers les neuf cellules Dans-la-population à Dans-la-population de la matrice des flux bruts.

Dans la section 4.2, nous avons vu que les augmentations nettes des strates sont attribuées aux cellules d'entrée alors que les diminutions nettes sont attribuées aux cellules de sortie en fonction des fractions des individus observées classées comme Occupés, Chômeurs et Inactifs au cours du mois t et du mois t-1 respectivement. Cette technique d'estimation des entrées et des sorties n'est valide que si les individus qui entrent dans la population étudiée et qui en sortent constituent un échantillon aléatoire des individus et, par conséquent, sont "identiques" aux individus qui demeurent à l'intérieur de la population en question. Les individus inclus dans l'échantillon qui sont classés comme Hors-de-la-population étudiée apparaissent dans l'échantillon accidentellement plutôt qu'intentionnellement; l'enquête sur la population

$$\sum_{u \neq A} (m_{u,A} - m_{A,u}) + \sum_{u \neq B} (m_{u,B} - m_{B,u}) = m_{0,A} - m_{A,0} + m_{B,A} - m_{A,B} + m_{C,A} - m_{A,C} + m_{0,B} - m_{B,0} + m_{B,C} - m_{C,B} + m_{0,C} - m_{C,0} + m_{C,A} - m_{A,C} + m_{C,B} - m_{B,C}$$

A noter que les mouvements entre les strates A et B s'annulent, mais que les termes indiquant les mouvements entre les strates A et C et les strates B et C demeurent dans la somme.

En général, les cellules d'entrée contiennent des termes supplémentaires de forme  $m_{v,u} - m_{u,v}$  pour chaque strate v qui a perdu des sujets alors que la strate u a gagné des sujets. De même, la cellule de sortie contient des termes supplémentaires de forme  $m_{x,y} - m_{y,x}$  pour chaque strate y qui a augmenté alors que la strate x a perdu des sujets.

Dans la cellule d'entrée, la quantité  $\sum_{u \neq s} (m_{u,s} - m_{s,u})$  pour chaque strate s qui a gagné des sujets entre les mois t-1 et t sera positive, bien que chaque terme distinct de la somme ne soit pas nécessairement positif. Si

$$\sum_{u \neq s} (m_{u,s} - m_{s,u}) > m_{0,s} - m_{s,0} \quad (12)$$

alors la contribution à la strate s est supérieure à l'entrée dans la strate s en provenance de l'extérieur de la population étudiée. Cet excédent provient des termes de forme  $m_{u,v} - m_{v,u}$  comme décrit ci-haut. C'est-à-dire que la surestimation est due à des mouvements entre les strates à l'intérieur de la population. On obtient un résultat semblable pour la cellule Dans-la-population à Hors-de-la-population de la matrice.

Les analystes de Statistique Canada signalent que la méthode qu'ils ont proposée pour le traitement des différences de poids entre les mois semble produire des surestimations des cellules d'entrée et de sortie de la matrice des flux bruts. Bien qu'ils soient basés sur des postulats simplificateurs, nos résultats indiquent une explication possible pour cette surestimation, qui peut être imputable à des mouvements à l'intérieur de la population étudiée.

est trouvée de la même façon. Ainsi, la matrice des flux bruts 2X2 se présen

te comme suit

Mois t

Dans-la-population      Hors-de-la-population

Mois t-1	Dans-la-population	Hors-de-la-population		
			0	
$\sum_{s=1}^S \{N_{t-1}^{s, \min} [0, \sum_{u \neq s} (m_{u,s-m,s,u})]\}$	$\sum_{s=1}^S \max [0, \sum_{u \neq s} (m_{u,s} - m_{s,u})]$			

$\sum_{s=1}^S N_{t-1}^{s, \min}$

$$\sum_{s=1}^S \{N_{t-1}^{s, \min} + \sum_{u \neq s} (m_{u,s} - m_{s,u})\}$$

Considérons la quantité qui se trouve dans la cellule entrée de cette matrice des flux bruts. Cette cellule devrait contenir l'augmentation nette de la population qui de la population qui provient de l'extérieur de la population étudiée,  $m_{0,s} - m_{s,0}$ , pour chaque strate qui a reçu des sujets de l'extérieur de la population. Ce que cette cellule contient est  $\sum_{u \neq s} (m_{u,s} - m_{s,u})$  pour chaque strate  $s$  qui a augmenté à la suite de mouvements entre les strates et en provenance de l'extérieur de la population étudiée. La somme  $\sum_{u \neq s} (m_{u,s} - m_{s,u})$  inclut la quantité  $m_{0,s} - m_{s,0}$  mais elle peut également contenir d'autres termes.

Par exemple, supposons que la population est constituée de trois strates appelées A, B et C. Si les strates A et B ont augmenté entre le mois t-1 et le mois t, et que la strate C a perdu des sujets, alors la cellule d'entrée con-

tient



$$N_s^{t-1} = N_s^{t-1} + \sum_{u \neq s} (m_{u,s} - m_{s,u}) \cdot \quad (7)$$

Les poids attribués aux individus dans la strate s dans les mois t-1 et t respectivement sont

$$W_s^{t-1} = N_s^{t-1} / r_s^{t-1}, t \text{ et } W_s^t = N_s^t / r_s^t \text{ et } W_s^{t-1} = N_s^{t-1} + \sum_{u \neq s} (m_{u,s} - m_{s,u}) / r_s^{t-1}, t. \quad (8)$$

Etant donné que nous attachons dans cette section au mouvement d'entrée dans la population étudiée et de sortie de celle-ci, il n'est pas nécessaire pour nous de diviser les membres de la population étudiée entre Occupés, Chômeurs et Inactifs. Ainsi, la matrice des flux bruts utilisée ici est une matrice  $2 \times 2$  formée par la sommation des trois premières lignes et des trois premières colonnes de la matrice des flux bruts  $4 \times 4$  utilisée dans la sous-section précédente.

L'entrée pour la strate s dans la cellule Dans-la-population à Dans-la-population est de

$$\min(W_s^{t-1}, W_s^t) r_s^{t-1}, t = \min[N_s^{t-1} / r_s^{t-1}, t, [N_s^{t-1} + \sum_{u \neq s} (m_{u,s} - m_{s,u})] / r_s^{t-1}, t]$$

$$= \min[N_s^{t-1}, N_s^{t-1} + \sum_{u \neq s} (m_{u,s} - m_{s,u})]$$

$$= N_s^{t-1} + \min[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})] \cdot \quad (9)$$

La valeur de la strate s dans la cellule Hors-de-la-population à Dans-la-population, ou entrée est

$$\max(0, W_s^{t-1} r_s^{t-1}, t = \max[0, \sum_{u \neq s} (m_{u,s} - m_{s,u}) / r_s^{t-1}, t]$$

$$= \max[0, \sum_{u \neq s} (m_{u,s} - m_{s,u})] \cdot \quad (10)$$

La valeur de la cellule Dans-la-population à Hors-de-la-population, ou sortie



taille des strates peut produire des estimations biaisées des cellules d'entrée et de sortie de la matrice des flux bruts.

#### 4.3 Effets des mouvements entre les strates

Les poids utilisés aux fins de l'estimation des flux bruts figurant dans l'expression (3) sont déterminés par le nombre de répondants des deux mois  $t-1$  et  $t$ , une quantité qui demeure constante pour les deux mois, et par la population des strates. La population d'une strate change si a) des individus y entrent venant de l'extérieur de la population étudiée, comme lorsque les gens atteignent leur quinzième année ou quittent les Forces armées à plein temps, b) des individus sortent de la population étudiée, comme lorsqu'ils entrent dans les Forces armées ou une institution, ou c) des individus de la population étudiée changent de strate. La présente sous-section décrit les effets de ces changements de la taille de la population sur les valeurs de la matrice des flux bruts.

Comme dans la sous-section précédente, nous supposons que la population étudiée est divisée en 5 strates. Encore une fois, les individus de l'échantillon sont tirés au hasard de chaque strate chaque mois, sont interviewés six mois consécutifs et sont ensuite retirés de l'échantillon. Soit  $r_{s,t-1}$ , comme auparavant, le nombre d'individus de la strate  $s$  qui sont interviewés dans les deux mois  $t-1$  et  $t$ . Ensuite, nous supposons qu'il y a  $N_{s,t-1}$  individus dans la strate  $s$  au cours du mois  $t-1$ . Indiquons les mouvements vers cette strate et hors de celle-ci entre les mois  $t-1$  et  $t$  par :

$m_{u,v}$  = nombre d'individus qui passent de  $u$  à  $v$ ,  $u \neq v$ , entre les entrevues pour les mois  $t-1$ , et  $t$  où  $u$  et  $v$  peuvent prendre les valeurs  $s = \text{strate } s \text{ pour } s = 1, 2, \dots, 5 \text{ et}$   
 $H = \text{hors de la population étudiée.}$

En employant cette notation, la population de la strate  $s$  dans le mois  $t$  est

A noter que chaque terme de la somme des neuf cellules de sujets dans la population active (les cellules qui indiquent les flux bruts entre Occupés, Chômeurs, et Inactifs) est le produit de la taille nette de la strate et de la fraction observée des sujets qui avaient les diverses classifications de la population active dans les mois  $t-1$  et  $t$ . La cellule Hors-de-la-population vers Occupés de la matrice des flux bruts contient la somme des termes de chaque strate qui a augmenté entre le mois  $t-1$  et le mois  $t$ . Chaque terme est le produit de l'augmentation nette de la taille de la strate et de la fraction des sujets de la strate qui se sont déclarés Occupés le mois  $t$ . Les cellules Hors-de-la-population à Chômeurs et à Inactifs contiennent les sommes de termes semblables sauf que l'augmentation nette de la taille de chaque strate est multipliée par la fraction des sujets de la strate qui étaient Chômeurs ou Inactifs respectivement le mois  $t$ . En d'autres mots, l'augmentation nette de la taille de chaque strate est attribuée proportionnellement aux trois cellules d'entrée de la matrice des flux bruts sur la base des fractions observées des Occupés, des Chômeurs et des Inactifs au cours du mois  $t$ . De même, la diminution nette de la taille de chaque strate qui a diminué entre les mois  $t-1$  et  $t$  est attribuée proportionnellement aux cellules de sortie de la matrice sur la base des fractions observées des Occupés, des Chômeurs et des Inactifs au cours du mois  $t-1$ .

Dans ce modèle, nous avons postulé que la seule façon pour des chiffres d'apparaître dans les cellules d'entrée et de sortie est une différence de poids. En pratique, un petit nombre d'individus qui entrent et sortent de la population étudiée apparaissent dans l'échantillon et les poids qui leur sont attribués sont ajoutés aux cellules d'entrée et de sortie appropriées. L'effet de ces individus sur les estimations est très petit.

Les fractions  $f_{s0}^s, f_{s1}^s, f_{s2}^s, f_{s3}^s, f_{s4}^s, f_{s5}^s, f_{s6}^s, f_{s7}^s, f_{s8}^s, f_{s9}^s$  sont estimées en utilisant les individus qui apparaissent dans l'échantillon au cours des deux mois. Presque tous les individus classés, par exemple comme  $H_0$ , ne pourraient pas être répondants dans les deux mois parce qu'ils n'ont pas été choisis comme sujet à cause du plan ou parce qu'ils ont déménagé. Ainsi, les gens qui n'auraient pu être répondants les deux mois sont représentées par des individus qui ont été répondants ces deux mois. Dans la mesure où ces groupes sont différents, l'allocation proportionnelle des augmentations et des diminutions nettes de la

$$= \max(0, N_t^s - N_{t-1}^s) n_{00}^s / r_{t-1, t}.$$

Les différences des individus tombant dans les cellules CO et IO contribueront également à la cellule HO. Ainsi, la contribution totale de la cellule HO de la strate est:

$$\max(0, N_t^s - N_{t-1}^s) \{ (n_{00}^s / r_{t-1, t}^s) + (n_{IO}^s / r_{t-1, t}^s) \}$$

$$= \max[0, N_t^s - N_{t-1}^s] n_{+0}^s / r_{t-1, t}^s$$

$$= \max(0, N_t^s - N_{t-1}^s) f_{+0}^s$$

(6)

où  $f_{+0}^s$  = fraction de tous les individus de la strate s, interviewés dans les deux mois t-1 et t, qui étaient occupés dans le mois t.

On obtient les totaux de toutes les cellules de la matrice des flux bruts de façon semblable. La matrice des flux bruts qui en résulte est la suivante:

Matrice des flux bruts - Mois t-1 à mois t

Mois t

	0	C	I	H
0	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$
C	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$
I	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$
H	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$	$\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$ $\sum_s \min(N, N_{t-1}^f) f_{+0}^s$

$$w_{s,t-1}^s = N_{s,t-1}^s / r_{s,t-1}^s \text{ et } w_s^t = N_s^t / r_{s,t-1}^t. \quad (3)$$

Aussi longtemps que ces mouvements entre les strates et la sélection des groupes constants sont des "processus aléatoires", ces poids représentent l'inverse de la probabilité qu'un individu à l'intérieur d'une strate soit interviewé dans les deux mois  $t-1$  et  $t$ . Étant donné que tous les individus à l'intérieur d'une strate ont le même poids un mois donné, on peut utiliser des agrégats pour chaque strate. Par conséquent, nous définissons :

$$n_{ij}^s = \text{nombre d'individus inclus dans l'échantillon de la strate } s \text{ classée dans la catégorie d'activité } i \text{ au mois } t-1 \text{ et } j \text{ la catégorie } j \text{ au mois } t \text{ pour } i, j = 0, C, I, H.$$

La méthodologie proposée par Statistique Canada exige que le minimum des poids des mois  $t-1$  et  $t$  de chaque individu soit ajouté à la cellule appropriée dans la matrice des flux bruts. Cette différence est ajoutée à la cellule d'entrée appropriée si le poids du mois  $t$  est supérieur au poids du mois  $t-1$  et à la cellule de sortie appropriée dans le cas contraire. Ainsi, par exemple, l'entrée de la strate  $s$  dans la cellule  $00$  (occupés à occupés) de la matrice des flux bruts est :

$$\min(w_{s,t-1}^s, w_s^t) n_{00}^s = \min [N_{s,t-1}^s / r_{s,t-1}^s, (N_s^t / r_{s,t-1}^t)] n_{00}^s$$

$$= \min (N_{s,t-1}^s, N_s^t) N_{00}^s / r_{s,t-1}^s$$

$$= \min (N_{s,t-1}^s, N_s^t) f_{00}^s$$

(4)

où  $f_{00}^s$  = fraction de tous les individus de la strate  $s$ , interviewés les deux mois  $t-1$  et  $t$ , qui étaient occupés ces deux mois.

La contribution de la strate  $s$  à la cellule  $00$  de la matrice des individus occupés dans le mois  $t$  est :

$$\max(0, w_{s,t-1}^s - w_s^t) n_{00}^s = \max [0, (N_{s,t-1}^s - N_s^t) / r_{s,t-1}^s] n_{00}^s \quad (5)$$

Étant donné que le plan de l'enquête sur la population active est plutôt complexe, nous commençons avec une série de postulats simplificateurs. Dans notre modèle, nous postulons que :

1. on choisit un échantillon stratifié à un seul degré

2. il n'y a pas d'erreur de réponse et

3. les non-réponses ne se produisent que parce que des individus au hasard changent de strates ou à cause de leur insertion ou de leur retrait de l'échantillon.

#### 4.2 Attribution des changements nets de population aux cellules d'entrée et de sortie

Supposons que la population étudiée est divisée en  $S$  strates identifiées par  $s = 1, 2, \dots, S$ . Soit

$N_k^s$  = taille de la population dans la strate  $s$  au mois  $k$ .

Chaque mois, un échantillon aléatoire simple est choisi de chaque strate pour l'enquête et les individus choisis sont interviewés six mois consécutifs avant d'être retirés de l'échantillon. Notre but est d'estimer les flux bruts entre le mois  $t-1$  et le mois  $t$ .

Aux fins de l'estimation des flux bruts, seuls les individus qui sont interviewés dans les deux mois  $t-1$  et  $t$  seront utilisés. Ceci exclut les individus qui sont insérés dans l'échantillon ou retirés de celui-ci et les personnes qui passent d'une strate à l'autre. Soit

$r_s^{t-1, t}$  = le nombre d'individus inclus dans l'échantillon de la strate  $s$  interviewée dans les mois  $t-1$  et  $t$ .

Chacun des  $r_s^{t-1, t}$  répondants de la strate  $s$  reçoit les poids suivants dans les mois  $t-1$  et  $t$  respectivement aux fins de l'estimation des flux bruts :



d'une étape à l'autre.

Des vérifications à Statistique Canada ont montré que les changements de cellules résultant de l'application de la remise à l'échelle proportionnelle itérative étaient petits en valeur absolue et en valeur relative et se situaient généralement à l'intérieur des limites de variation de l'échantillon associées aux cellules. Ceci laisse croire que la correction pour la convergence ne biaise pas sérieusement les estimations des flux bruts.

### 3.5 Correction du biais d'erreur de classification

La méthode proposée par Statistique Canada pour l'estimation des flux bruts comprend également une étape de correction du biais d'erreur de classification. Il s'agit du biais qui résulte de la mauvaise attribution du statut d'activité d'un individu. Une technique mise au point par Fred Wong (1983) à Statistique Canada utilise les données d'une nouvelle entrevue pour corriger le biais d'erreur de classification.

## **4. CONSÉQUENCE DE LA MÉTHODE PROPOSÉE PAR STATISTIQUE CANADA**

### 4.1 Les modèles des flux bruts

Chaque étape de la méthode proposée par Statistique Canada décrite dans la section précédente constitue une tentative logique de corriger des problèmes concernant la production d'estimations valides des flux bruts. On ne connaît pas exactement, cependant, l'effet des diverses corrections sur les estimations finales de la matrice des flux bruts. Pour mieux comprendre la proposition de Statistique Canada de traiter les différences de poids comme étant attribuables aux entrées dans la population étudiée et aux sorties de celle-ci, nous exposerons dans la présente section un modèle du processus de flux bruts. Notre discussion se centrera sur les valeurs des cellules d'entrée et de sortie de la matrice des flux bruts estimés, étant donné que les problèmes affectant la méthode proposée par Statistique Canada semblent se produire dans ces cellules.

proportionnellement de façon que les entrées et les sorties totales indiquées dans la matrice des flux bruts soient égales aux estimations du recensement des entrées et des sorties respectivement.

Soit I l'estimation indépendante du recensement des entrées dans la population étudiée et F les estimations du recensement des sorties de cette population. Appelons la somme des entrées estimées  $X_{H+} = X_{H0} + X_{HC} + X_{HI}$  et la somme des sorties estimées  $X_{+H} = X_{0H} + X_{CH} + X_{IH}$ . Les entrées ajustées proportionnellement sont

$$Y_{Hj} = X_{Hj} I / X_{H+} \quad \text{pour } j = 0, C, I. \quad (1)$$

Les sorties ajustées proportionnellement sont

$$Y_{jH} = X_{jH} F / X_{+H} \quad \text{pour } j = 0, C, I. \quad (2)$$

### 3.4 Convergence des estimations des flux bruts avec les totalisations

mensuelles

Statistique Canada aimerait que les estimations des flux bruts convergent avec les estimations mensuelles publiées des taux d'activité totaux. Ainsi, les totaux des lignes de la matrice des flux bruts doivent être les estimations du taux d'activité du mois t-1 et les totaux des colonnes doivent être les estimations ponctuelles du mois t. Les totaux marginaux de la matrice des flux bruts établis comme on l'a décrit ci-haut ne convergent pas avec les totaux mensuels de la population active.

Statistique Canada a l'intention d'utiliser la méthode de remise à l'échelle proportionnelle itérative, originellement proposée par Deming et Stephan (1940) et décrite en détail par Bishop, Fienberg et Holland (1975), pour corriger la matrice des flux bruts en fonction des totaux mensuels de la population active. Lorsqu'elle est utilisée pour corriger la matrice des flux bruts, la remise à l'échelle proportionnelle itérative successivement 1) ajuste les lignes de la matrice à la somme des estimations t-1 et ensuite 2) ajuste les colonnes à la somme des estimations du mois t. On répète les étapes 1) et 2) jusqu'à ce que les données de la matrice ne changent pas

différence,  $W_{t,i} - W_{t-1,i}$ , est ajoutée à la cellule HO, étant donné qu'on estimerait que ces cinq personnes se trouvaient en dehors de la population étudiée au cours du mois  $t-1$  et se sont ajoutées à la population comme personnes occupées au cours du mois  $t$ .

Si, par ailleurs, l'individu est occupé les deux mois, mais que  $W_{t-1,i} = 305$  et que  $W_{t,i} = 300$ , alors on ajoute encore une fois le 300 à la cellule OO mais le poids supplémentaire de 5 est ajouté à la cellule OH. Ici, la différence entre les poids représente 5 personnes qui étaient occupées au cours du mois  $t-1$  et qui sont sorties (Hors) de la population étudiée au cours du mois  $t$ .

Un individu classé comme Hors de la population étudiée dans le mois  $t-1$  et qui est ensuite, disons, occupé dans le mois  $t$ , aura un  $W_{t-1,i} = 0$ . Si  $W_{t,i} = 300$  alors on ajoute 300 à la cellule HO. Les individus classés comme Hors-de-la population étudiée au cours du mois  $t$  sont traités de la même façon, en ajoutant  $W_{t-1,i}$  à la cellule appropriée dans la dernière colonne de la matrice des flux bruts.

Parce que les personnes Hors de la population étudiée se voient attribuer un poids zéro, une personne classée ainsi dans les deux mois  $t-1$  et  $t$  auraient donc  $W_{t-1,i} = 0$  et  $W_{t,i} = 0$ . Par conséquent,  $X_{HH}$ , la valeur de la cellule HH de la matrice des flux bruts doit toujours être de zéro.

### 3.3 Correction des cellules d'entrée et de sortie

L'addition des différences des poids aux cellules d'entrée et de sortie de la matrice des flux bruts fournit une méthode de traitement des changements des poids basés sur l'échantillon d'un mois à l'autre et fournit des estimations des entrées dans la population étudiée et des sorties de celle-ci. Des estimations indépendantes des entrées et des sorties, disponibles à partir des données du recensement, laissent croire que cette méthode surestime la valeur réelle des mouvements dans et hors de la population étudiée. Ainsi, Statistique Canada a l'intention d'ajuster les  $X_{HO}$ ,  $X_{HC}$ ,  $X_{HI}$ ,  $X_{OH}$ ,  $X_{CH}$ , et  $X_{IH}$  dans la matrice des flux bruts. Ces cellules seront corrigées

non-réponse d'un mois à l'autre. Statistique Canada propose de répondre les fichiers des individus qui ont répondu dans les deux mois  $t-1$  et  $t$  pour com-

prendre pour cette non-réponse supplémentaire.

Lorsque la répondération sera terminée, Statistique Canada aura en main un fichier de données uniques comprenant des renseignements provenant de toutes les personnes qui ont répondu deux mois consécutifs. Ce fichier contiendra des renseignements géographiques et démographiques pour chaque individu ainsi que le statut d'activité de l'individu et la pondération qui lui est attribuée pour les mois  $t-1$  et  $t$ .

### 3.2 Différences de poids

Comme nous l'avons noté dans la section 2.3, le poids attribué à un individu de la non-réponse et des changements dans la taille de la population étudiée. Même lorsque le facteur de correction pour la non-réponse est calculé sur les données d'un mois à l'autre, il peut encore arriver, pour un individu quelconque  $i$ , que  $W_{t-1,i} \neq W_{t,i}$ . Si l'on veut utiliser les données pour estimer les flux bruts, il est essentiel d'avoir une méthode pour traiter cette différence de pondération.

Statistique Canada propose de résoudre ce dilemme en postulant que les différences entre les deux poids ne se produisent que comme résultat des entrées et des sorties de la population étudiée. Ainsi, les différences des poids sont ajoutées à la cellule appropriée soit dans la dernière rangée ou la dernière colonne de la matrice des flux bruts. Cette procédure est fortement dépendante de l'interprétation des poids suggérés à la section 2.3, c'est-à-dire que l'individu  $i$  de l'échantillon représente  $W_{t,i}$  personnes de la population au cours du mois  $t$ .

Pour illustrer cette procédure, supposons qu'un individu soit classé comme occupé dans les mois  $t-1$  et  $t$ , mais que  $W_{t-1,i} = 300$  et  $W_{t,i} = 305$ . Le poids minimum, 300, est ajouté à la cellule 00 du tableau des flux bruts. Cette



Statut d'activité pour le mois t

O C I H

Statut d'activité pour le mois t-1	O	C	I	H
O	$X_{OO}$	$X_{OC}$	$X_{OI}$	$X_{OH}$
C	$X_{CO}$	$X_{CC}$	$X_{CI}$	$X_{CH}$
I	$X_{IO}$	$X_{IC}$	$X_{II}$	$X_{IH}$
H	$X_{HO}$	$X_{HC}$	$X_{HI}$	$X_{HH}$

où :

0 = Occupé

C = Chômeur

I = Inactif

H = Hors de la population étudiée, et

$X_{ij}$  =

nombre estimé de personnes ayant le statut d'activité i dans le mois t-1 et le statut j dans le mois t.

On peut utiliser les dossiers finals de l'enquête sur la population active pour deux mois consécutifs pour obtenir les données d'estimation des flux bruts de la matrice 4x4. Pour utiliser ces données pour la production des estimations des flux bruts, Statistique Canada doit apparier les dossiers individuels extraits des deux fichiers mensuels consécutifs à l'aide des numéros d'identification personnels attribués aux individus enquêtés pour la durée de leur passage dans l'enquête.

Un individu présent dans le fichier des données dans un mois donné peut être absent du fichier dans un autre mois à cause de son insertion dans l'échantillon ou de son retrait de celui-ci ou pour avoir déménagé, avoir été absent de la maison ou avoir refusé de répondre. La pondération de l'échantillon décrite dans la section 2.4 comprend une correction pour la non-réponse chaque mois. Dans le traitement des flux bruts, nous devons également considérer la



n'est pas nécessairement vrai que  $W_{t-1,i} = W_{t,i}$ .

## 2.4 Structure longitudinale et estimation des flux bruts

Bien que l'objectif principal de l'enquête sur la population active soit de produire des estimations ponctuelles de l'activité sur le marché du travail, la technique du panel de l'enquête produit une base de données longitudinales dans laquelle environ les cinq sixièmes des ménages enquêtés un mois donné se retrouvent dans l'échantillon le mois suivant. Évidemment, on rejoint moins des cinq sixièmes des individus ou des ménages de l'échantillon au cours de deux mois consécutifs à cause de la non-réponse et des déménagements. Cependant, Statistique Canada s'intéresse à la possibilité d'utiliser les renseignements des individus qui ont répondu deux mois consécutifs pour produire des estimations des flux bruts entre les catégories de la population active.

Les estimations des flux bruts servent à répondre à des questions telles que a) quelle proportion de l'augmentation du chômage est due à des pertes d'emploi et quelle proportion est due à des personnes non auparavant actives comptant à se chercher un emploi? ou b) combien de personnes en chômage se découragent et quittent la population active?

Nous discutons du problème de l'estimation des flux bruts entre les catégories de la population active dans les deux prochaines sections.

## 3. MÉTHODE PROPOSÉE PAR STATISTIQUE CANADA POUR L'ESTIMATION DES FLUX BRUTS

Dans cette section, nous décrivons la procédure à plusieurs degrés d'estimation des flux bruts mise au point par Statistique Canada (voir Macredie et Vevers, 1977; Wong, 1983). Notre description de la procédure comprend diverses interprétations de l'effet des divers degrés.

### 3.1 Données nécessaires à l'estimation des flux bruts

Statistique Canada a proposé d'estimer les flux bruts à l'aide d'une matrice

population active sont les ménages. On trouvera une description détaillée du

plan d'échantillonnage de l'enquête dans Méthodologie de l'enquête sur la population active, 1976, Statistique Canada (1977).

Les ménages choisis pour l'enquête sur la population active font partie de l'enquête pendant six mois consécutifs et sont ensuite retirés de l'échantillon. Par exemple, les ménages intégrés à l'échantillon en janvier sont interviewés six mois de suite, et ensuite retirés de l'échantillon après l'entrevue de juin. Chaque groupe de ménages ainsi intégré à l'échantillon et retiré ensuite constitue un panel. Chaque mois, l'échantillon de l'enquête sur la population active comprend des sujets de six groupes constants différents.

## 2.3 Pondération basée sur l'échantillon

Les données ponctuelles, les renseignements d'un mois donné pour tous les sujets dans les six panels interviewés au cours de ce mois, sont utilisées pour produire les estimations mensuelles de l'activité sur le marché du travail. Les estimations mensuelles sont des moyennes pondérées des valeurs pour chaque individu de l'échantillon. On utilise une moyenne pondérée parce que chaque sujet est considéré comme "représentant" un certain nombre de personnes dans la population étudiée. Le poids attribué au dossier d'un individu correspond au nombre de personnes de la population que cet individu de l'échantillon représente.

Soit  $W_{t,i}$  le poids attribué à l'individu  $i$  au mois  $t$ . Si l'individu  $i$  est classé comme à l'extérieur de la population étudiée au cours du mois  $t$ , alors  $W_{t,i} = 0$ . Sinon, les poids attribués sont fixés par la probabilité de sélection de la grappe, la probabilité de sélection du ménage à l'intérieur de la grappe, la non-réponse à l'intérieur du mois, les facteurs rural/urbain, les corrections de sous-échantillonnage pour les zones dont la croissance est rapide et l'ajustement des rapports pour les facteurs province, âge, sexe.

Le poids attribué à un individu peut changer d'un mois à l'autre à cause du fait que chaque mois, on remplace un sixième de l'échantillon, à cause également de la non-réponse et, dans une moindre mesure, à cause des changements de la taille de la population étudiée. Ainsi, pour un individu donné,  $i$ , il

## 2. DESCRIPTION DE L'ENQUÊTE SUR LA POPULATION ACTIVE

### 2.1 Champ d'application de l'enquête

Environ 56,000 ménages, choisis dans les dix provinces canadiennes, sont inclus chaque mois dans l'échantillon de l'enquête sur la population active. Un questionnaire est rempli pour chaque membre des ménages de l'échantillon âgé de 15 ans et plus, ne faisant pas partie des forces armées et ne vivant pas en institution. Les questions de l'enquête concernent principalement les activités du sujet reliées au travail au cours de la semaine de référence, qui est la semaine précédant la semaine de l'enquête et qui contient généralement le 15<sup>e</sup> jour du mois. Les réponses aux questions de l'enquête sont utilisées pour classer les sujets comme occupés, chômeurs ou inactifs. Pour une discussion sur la classification de l'activité, voir Guide d'utilisation des données de l'enquête sur la population active, Statistique Canada (1979).

### 2.2 Conception de l'enquête

L'enquête sur la population active a été conçue pour permettre l'estimation des niveaux et des taux de l'emploi et du chômage pour chacune des dix provinces individuellement. Ainsi, excepté en ce qui concerne les exigences pour la taille totale de l'échantillon, les échantillons de chaque province sont indépendants.

Les régions économiques (RE), zones de structure économique semblables, forment la strate fondamentale à l'intérieur des provinces. Les RE sont divisées en unités autoreprésentatives (UAR) et unités non autoreprésentatives (UNAR). Les UAR sont des centres urbains importants et les UNAR sont généralement composées d'un petit centre urbain et d'une zone rurale. L'échantillonnage est effectué séparément pour les UAR et les UNAR.

L'échantillon des UAR est un échantillon stratifié à deux degrés. L'échantillonnage des UNAR est effectué selon un schéma d'échantillonnage stratifié à plusieurs degrés. En plus des UAR et des UNAR, certaines unités d'échantillonnage sont tirées d'un univers des immeubles d'appartements et d'un univers des zones spéciales. Les unités d'échantillonnage finales de l'enquête sur la

multiples des sujets fournissent une base de données longitudinales additionnelles qui pourrait être exploitée pour fournir des estimations des changements dans le temps pour des coûts supplémentaires minimes (voir, par exemple, Kalachek, 1979, et Fienberg et Tanur, 1983).

On a fait quelques tentatives d'utiliser les données longitudinales tirées des enquêtes par panel. Par exemple, les données longitudinales de la Current Population Survey sont utilisées depuis 1948 pour produire des tableaux montrant les mouvements bruts des individus entre les diverses catégories de la population active d'un mois à l'autre. Bien que ces tableaux soient produits chaque mois, ils n'ont pas été publiés depuis 1952 à cause de problèmes statistiques. Smith et Vanski (1979) discutent la production de données des variations brutes en utilisant la structure longitudinale de la Current Population Survey.

Récemment, Statistique Canada a commencé une étude sur les utilisations possibles des données longitudinales disponibles comme sous-produits de l'enquête sur la population active. Cux aussi souhaiteraient trouver une méthode de production d'estimations fiables des mouvements bruts entre les catégories de la population active. Dans le présent document, nous discutons les méthodes à l'étude à Statistique Canada pour la production de ces données des variations brutes.

Dans la section 2, nous donnons une brève description du champ d'application et de la conception de l'enquête sur la population active, et nous décrivons la structure des données produites. À la section 3, nous décrivons les méthodes proposées d'estimation des flux bruts mises au point par Statistique Canada, qui nécessitent l'utilisation de poids fonction de l'échantillon, de corrections pour les entrées dans la population étudiée et les sorties de celle-ci, de corrections pour l'homogénéité et de corrections pour le biais causé par les erreurs de classification. En mettant au point certains modèles simples pour le processus des flux bruts, nous analysons dans la section 4 les conséquences de la méthode proposée par Statistique Canada. Finalement, dans la section 5, nous décrivons certains travaux sur le traitement de la non-réponse dans les estimations des flux bruts.



# L'ESTIMATION DES FLUX BRUTS MENSUELS DE L'ACTIVITÉ SUR LE MARCHÉ DU TRAVAIL<sup>1</sup>

Stephen E. Fienberg et Elizabeth A. Stasny<sup>2</sup>

L'enquête sur la population active au Canada est une enquête mensuelle sur les ménages effectuée chaque mois dans le but de fournir des estimations ponctuelles du nombre de personnes occupées, en chômage ou inactives. L'enquête utilise une technique de groupe constant avec renouvellement dans laquelle tous les individus d'un ménage inclus dans l'échantillon sont interviewés chaque mois, six mois de suite. Dans le passé, cette structure longitudinale a été peu exploitée, bien qu'on ait montré beaucoup d'intérêt pour les flux bruts d'un mois à l'autre (transitions) entre ces diverses catégories d'activité. Dans le présent texte, nous discutons des méthodes de production des estimations des flux bruts à l'étude à Statistique Canada, mais à partir du point de vue de la production d'un modèle.

## 1. INTRODUCTION

L'enquête sur la population active au Canada est une enquête mensuelle sur les ménages utilisée pour produire des estimations ponctuelles, ou transversales, de l'activité sur le marché du travail. Cette enquête, cependant, comme la Current Population Survey aux États-Unis et beaucoup d'autres enquêtes par échantillonage d'envergne, utilise la technique du panel où les sujets sont interviewés plusieurs fois avant d'être retirés de l'échantillon. Bien que le principal but de l'enquête soit l'obtention d'estimations ponctuelles, on a depuis longtemps constaté que les renseignements provenant de ces interviews

<sup>1</sup> Cette recherche a été financée en partie par un contrat avec Statistique Canada. Les auteurs désirent remercier Murray Lawes, Larry Swain et Richard Vevers pour leurs explications de la méthodologie de l'enquête sur la population active et de ses données ainsi que l'éditeur et un évaluateur pour leurs commentaires et suggestions utiles.

<sup>2</sup> Stephen E. Fienberg et Elizabeth A. Stasny, Carnegie-Mellon University, Pittsburgh, Pennsylvania, PA 15213.



Incidence du cancer  
1969-1978  
Cancers primitifs multiples de chaque siège  
(Canada à l'exclusion de l'Ontario)

Cancers primitifs multiples		Siège du cancer ICDA			
% ayant le même chiffre du code ICDA		Nombre	% d'in- cidence	% dans le même registre	
180	Col de l'utérus	135	1.4	60.9	S/O
181	Chorionéplionie	1	1.0	0.0	S/O
182	Autre: de l'utérus	149	1.1	73.8	91.3
183	Ovaire, etc.	126	1.5	80.2	97.6
184	Org. gén. fem., autres	31	1.5	71.0	83.9
185	Prostate	64	1.6	69.2	5/0
186	Testicule	27	1.5	44.4	5/0
187	Org. gén. masc., autres	9	1.4	88.9	100.0
188	Vessie	274	1.6	77.7	5/0
189	Org. urin. autres et sans autre	119	1.5	72.3	79.8
190	Oeil	13	1.1	46.2	5/0
191	Cerveau	93	1.6	46.2	5/0
192	Autre système nerveux	6	0.4	33.3	66.7
193	Glande thyroïde	49	1.5	55.1	5/0
194	Autres glandes endocrines	5	0.6	80.0	100.0
195	Stgès mal définis	4	0.6	100.0	100.0
196	Ganglions lymphatiques sec. et non	5	0.3	100.0	100.0
197	Secondaires; Resp. et Digestif	9	0.3	88.9	77.8
198	Autres secondaires	1	0.1	100.0	100.0
199	Sans précision du siège	12	0.3	75.0	83.3
200	Lymphosarcome, etc.	68	1.1	60.3	97.1
201	Maladie de Hodgkin	85	2.2	55.3	5/0
202	Autre, du tissu lymphoïde	22	0.8	90.9	72.7
203	Mylome multiple	45	1.3	64.4	5/0
204	Leucémie lymphatique	64	1.5	62.5	84.4
205	Leucémie de la moelle	32	1.0	53.1	90.6
206	Leucémie monocyttaire	4	1.0	75.0	100.0
207	Leucémie, autre et non spécifiée	11	0.7	81.8	90.9
208	Polyglobulie essentielle	1	0.2	100.0	5/0
209	Myléofibrose	1	0.4	100.0	5/0

Tableau 7

**Incidence du cancer  
1969-1978  
Cancers primitifs multiples de chaque siège  
(Canada à l'exclusion de l'Ontario)**

Siège du cancer ICDA		Nombre	% d'in- cidence	% dans le même registre	% ayant le même qu chiffre du code ICDA
Cancers primitifs multiples					
Total des sièges (excepté la peau, 173)					
140	Lèvre	87	1.4	90.8	83.9
141	Langue	30	1.6	63.3	63.3
142	Glande salivaire	8	0.6	12.5	75.0
143	Gencive	8	1.5	100.0	62.5
144	Plancher buccal	18	1.8	94.4	62.5
145	Bouche, autre et non spécifique	16	1.4	75.0	62.5
146	Arrière-gorge	14	1.0	92.9	78.6
147	Rhinopharynx	8	1.1	62.5	5/0
148	Laryngopharynx	3	0.5	100.0	33.3
149	Pharynx, non spécifique	3	1.3	66.7	5/0
150	Oesophage	37	1.1	75.7	5/0
151	Estomac	178	1.0	71.9	74.2
152	Petit intestin	12	1.2	91.7	91.7
153	Gros intestin, à l'exclusion du rectum	623	1.9	82.3	44.0
154	Rectum	256	1.4	77.0	74.6
155	Foie	11	0.6	63.6	90.9
156	Vésicule biliaire	22	0.7	77.3	77.3
157	Pancréas	96	1.0	70.8	54.2
158	Péritoine	6	0.7	100.0	100.0
159	Organes digestifs non spécifiés	1	0.3	100.0	5/0
160	Nez, etc.	5	0.7	80.0	80.0
161	Larynx	94	2.0	77.7	57.4
162	Trachée, bronches, poumon	795	1.9	75.0	99.5
163	Org. resp. autres et sans autre	7	0.7	85.7	85.7
170	Os	28	2.0	78.6	82.1
171	Tissus conjonctifs	32	1.3	81.3	75.0
172	Mélanome de la peau	31	1.4	78.7	54.1
174	Sein	1,971	3.6	84.4	5/0

Tableau 6

Mortalité du cancer  
Disponibilité et intégralité des données élémentaires

Canada (à l'exclusion de l'Ontario) et les provinces comptant les pourcentages  
les plus élevés et les moins élevés d'intégralité des données

Données élémentaires	Année de décès	Canada	Pourcentage le plus élevé	Pourcentage le moins élevé
DATE DE NAISSANCE				
Jour	1969 - 1973 1974 - 1978 1969 - 1978	96.6 97.8 97.2	99.4 99.7 99.6	0.0 43.9 23.0
Mois	1969 - 1973 1974 - 1978 1969 - 1978	97.0 98.1 97.6	99.7 99.9 99.8	0.0 44.2 23.2
Année	1969 - 1973 1974 - 1978 1969 - 1978	100.0 99.9 99.9	100.0 100.0 100.0	99.2 99.6 99.4
Date de naissance complète	1969 - 1973 1974 - 1978 1969 - 1978	96.6 97.8 97.2	99.4 99.7 99.6	0.0 43.9 23.0
AGE	1969 - 1973 1974 - 1978 1969 - 1978	100.0 99.9 100.0	100.0 100.0 100.0	100.0 99.6 99.8
LIEU DE NAISSANCE (Pays ou province)	1969 - 1973 1974 - 1978 1969 - 1978	98.3 52.2 73.9	100.0 99.9 99.9	98.9 0.0 46.2
RÉSIDENCE Division du recen- sement	1969 - 1973 1974 - 1978 1969 - 1978	99.8 99.7 99.8	100.0 100.0 100.0	99.6 99.8 99.7
Subdivision du recensement	1969 - 1973 1974 - 1978 1969 - 1978	92.2 99.7 96.2	99.9 99.5 99.7	51.9 99.4 77.1

Tableau 5

**Incidence du cancer**  
**Disponibilité et intégralité des données élémentaires**  
 Canada (à l'exclusion de l'Ontario) et les provinces comptant les pourcentages  
 les plus élevés et les moins élevés d'intégralité des données

DONNÉES ÉLÉMENTAIRES	ANNÉE DE DIAGNOSTIC	CANADA	POURCENTAGE LE PLUS ÉLEVÉ	POURCENTAGE LE MOINS ÉLEVÉ
DATE DE MAISSANCE	Jour	1969 - 1973 1974 - 1978 1969 - 1978	98.8 99.8 99.3	0.0 4.0 2.2
	Mois	1969 - 1973 1974 - 1978 1969 - 1978	98.8 99.8 99.4	0.0 4.1 2.2
	Année	1969 - 1973 1964 - 1978 1969 - 1978	100.0 100.0 100.0	15.9 21.9 19.1
	Date de naissance complète	1969 - 1973 1974 - 1978 1969 - 1978	98.7 99.8 99.3	0.0 4.0 2.2
AGE	1969 - 1973 1974 - 1978 1969 - 1978	99.4 100.0 99.7	100.0 100.0 100.0	98.6 100.0 99.4
LIEU DE MAISSANCE (Pays ou province)	1969 - 1973 1974 - 1978 1969 - 1978	15.2 24.4 20.2	19.8 71.1 46.0	0.0 0.0 0.0
RÉSIDENCE Division du recen- sement	1969 - 1973 1974 - 1978 1969 - 1978	89.6 89.6 89.6	100.0 100.0 100.0	4.1 82.8 49.0
Subdivision du recensement	1969 - 1973 1974 - 1978 1969 - 1978	16.6 32.3 25.2	43.2 76.4 61.6	0.0 0.0 0.0

Tableau 4

Appariement des dossiers incidence aux dossiers mortalité

Classification des causes de décès pour un échantillon de cas de cancer du poumon  
tirés du fichier incidence du cancer (Distribution de pourcentage)

Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux d'appariement les plus élevés ou les moins élevés de cas de cancer du poumon parmi les appariements

ICDA-8	CLASSIFICATION SELON LE FICHIER MORTALITÉ			CANADA	TAUX LE + ÉLEVÉ	TAUX LE - ÉLEVÉ
162.1	(a) POU MON; primitif	91.4	92.0	100.0	85.3	
160-163	(b) AUTRES PARTIES DE L'APPAREIL RESPIRATOIRE; primitif	0.3	0.3	100.0	0.0	
197.0-197.3	(c) APPAREIL RESPIRATOIRE; secondaire	0.3	4.3	2.4		
2. AUTRES CANCERS		0.3	0.3	9.8		
174	(a) SEIN	0.3	0.0	0.0		
200-209	(b) SYSTÈME LYMPHATIQUE ET HÉMATOPOÏÉTIQUE; primitif	0.3	0.0	0.0		
196	Secondaire	-	0.0	0.0		
195, 199	(c) AUTRE SIÈGE PRIMITIF SPÉCIFIÉ (d) SIÈGE MAL DÉFINI OU NON DÉFINI	3.1 0.6	9.8 0.0	4.9		
3. PAS DE CANCER		3.7	100.0	100.0		
TOTAL		100.0	100.0	100.0		

TAILLE DE L'ÉCHANTILLON POUR LES APPARIEMENTS	350	45	41
TAUX DE RÉUSSITE DES APPARIEMENTS (%)	69.4	80.4	73.2



Tableau 3

Appariement des dossiers mortalité aux dossiers incidence

Classification des maladies selon le fichier mortalité cancer pour un échantillon de décès dus au cancer (Distribution de pourcentage)

Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux d'appariement les plus élevés ou les moins élevés de cas de cancer du poumon parmi les appariements

ICDA-8	CLASSIFICATION SELON LE FICHIER MORTALITÉ			CANADA	TAUX LE + ÉLEVÉ	TAUX LE - ÉLEVÉ
162.1	(a) POU MON; primitif	92.5	100.0	83.7	74.4	4.7
160-163	(b) AUTRES PARTIES DE L'APPAREIL RESPIRATOIRE; primitif	1.0		4.7		4.7
197.0-197.3	(c) APPAREIL RESPIRATOIRE; secondaire	1.7		4.7		4.7
2. AUTRES CANCERS						
174	(a) SEIN	4.8		16.3	2.3	0.0
200-209	(b) SYSTÈME LYMPHATIQUE ET HÉMATOPOÏÉTIQUE; primitif	0.7				0.0
196	Secondaire	1.0		4.6		9.3
195, 199	(c) AUTRE SIÈGE PRIMITIF SPÉCIFIÉ	1.9				0.0
	(d) SIÈGE MAL DÉFINI OU NON DÉFINI	0.5				100.0
TOTAL		100.0	100.0	100.0	100.0	100.0

TAILLE DE L'ÉCHANTILLON POUR LES APPARIEMENTS	415	54	43
TAUX DE RÉUSSITE DES APPARIEMENTS (%)	82.3	96.4	76.8

Tableau 2  
Ratios mortalité-incidence du cancer

Décès au cours de la période (mortalité) en pourcentage des  
nouveaux cas enregistrés (incidence)  
Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux  
les plus élevés ou les moins élevés.

Siège du cancer		Principaux sièges selon le groupe d'âge et le sexe													
		Âge		1969 - 1973							1974 - 1978				
				Canada		taux le plus élevé		taux le moins élevé			Canada		taux le plus élevé		taux le moins élevé
Sein (174)	Total	40	47	31	35	31	26	30*	43	55	46	37	46	31	
	0-24	14	13	-	31	26	30*	43	55	46	37	46	31		
	25-44	28	33	44	35	31	26	30*	43	55	46	37	46	31	
	45-64	37	62	62	35	31	26	30*	43	55	46	37	46	31	
	65+	50	50	50	35	31	26	30*	43	55	46	37	46	31	
Colon (153)	Total	73	94	50	55	41	43	55	74	67	80	33*	55	34	
	0-24	15	20	-	27	41	55	74	67	80	33*	55	34		
	25-44	48	49	41	43	55	74	67	80	33*	55	34			
	45-64	63	77	55	55	74	67	80	33*	55	34				
	65+	83	114	55	55	74	67	80	33*	55	34				
Utérus (182)	Total	27	38	17	20*	7	11	15	15	26	48	27	15	9	
	0-24	33	40*	20*	20*	7	11	15	15	26	48	27	15	9	
	25-44	13	19	9	20*	7	11	15	15	26	48	27	15	9	
	45-64	16	25	9	20*	7	11	15	15	26	48	27	15	9	
	65+	49	64	33	33	33	33	33	33	33	33	33	33	33	
Coi de l'utérus (180)	Total	39	45	29	32	5	15	25	49	65	78	100*	55	106	
	0-24	12	20*	17*	12	5	15	25	49	65	78	100*	55	106	
	25-44	21	17	12	12	5	15	25	49	65	78	100*	55	106	
	45-64	39	54	32	32	5	15	25	49	65	78	100*	55	106	
	65+	66	64	51	51	53	53	53	53	53	53	53	53	53	
Ovaire (183)	Total	69	79	54	54	69	28	38	72	98	64	45	39	58	
	0-24	20	33*	-	26	28	38	72	98	64	45	39	58		
	25-44	43	41	26	52	64	95	53	53	53	53	53	53		
	45-64	68	74	26	52	64	95	53	53	53	53	53	53		
	65+	87	105	77	77	95	53	53	53	53	53	53	53		
Tous les cancers (140-209) à l'exclusion du cancer de la peau (173)	Total	57	67	46	46	53	37	44	44	64	45	28	22	39	
	0-24	46	49	43	43	37	44	44	44	64	45	28	22	39	
	25-44	33	39	27	27	27	44	44	44	64	45	28	22	39	
	45-64	48	56	38	38	44	44	44	44	64	45	28	22	39	
	65+	75	94	58	58	69	44	44	44	44	44	44	44	44	

- Soit la mortalité, soit l'incidence soit les deux sont nulles.  
\* Taux basés sur moins de 10 cas tant pour la mortalité que pour l'incidence

**Tableau 1**  
**Ratios mortalité-incidence du cancer**  
**Décès au cours de la période (mortalité) en pourcentage des**  
**nouveaux cas enregistrés (incidence)**  
**Canada (à l'exclusion de l'Ontario) et les provinces comptant les taux**  
**les plus élevés ou les moins élevés.**

Hommes		Principaux sièges selon le groupe d'âge et le sexe									
Siège du cancer	Âge	1969 - 1973					1974 - 1978				
		Canada		taux le plus élevé		taux le moins élevé	Canada		taux le plus élevé		taux le moins élevé
Poumon (162)	Total	96	110	83	95	80	101	86	56	56	80
	0-24	27	37*	-	70	100*	-	-	-	-	100*
	25-44	89	96	83	85	82	75	87	45	-	82
	45-64	94	106	81	88	78	86	61	56	56	78
	65+	100	118	83	101	82	118	100	58	58	100
Prostate (185)	Total	44	55	35	39	33	54	86	56	56	33
	0-24	71*	133	-	71*	100*	-	-	-	-	-
	25-44	39	54	-	22	-	-	-	-	-	-
	45-64	27	32	33	24	20	30	87	45	45	20
	65+	48	62	35	43	36	60	61	56	56	36
Colon (153)	Total	76	101	54	67	56	86	86	56	56	56
	0-24	56	60	-	40	-	-	-	-	-	-
	25-44	61	76	40	53	45	87	87	45	45	87
	45-64	66	86	41	58	56	61	61	56	56	61
	65+	84	115	61	73	58	100	100	58	58	100
Vessie (188)	Total	35	42	26	28	23	35	35	23	23	23
	0-24	20	25*	33*	3	-	-	-	-	-	-
	25-44	7	9	10	6	-	-	-	-	-	-
	45-64	23	30	16	18	13	22	45	13	13	22
	65+	44	53	32	36	29	45	45	29	29	45
Estomac (151)	Total	104	124	82	90	80	112	112	80	80	80
	0-24	29*	-	-	86*	-	-	-	-	-	-
	25-44	82	102	54	74	77	40	40	77	77	40
	45-64	93	102	79	80	80	89	89	80	80	89
	65+	112	144	86	96	80	130	130	80	80	130
Tous les cancers à l'exclusion du cancer de la peau (173)	Total	69	83	56	64	56	77	77	56	56	56
	0-24	58	64	45	48	56	49	49	56	56	49
	25-44	51	62	39	45	37	55	55	37	37	55
	45-64	65	78	53	60	50	65	65	50	50	65
	65+	74	94	59	69	60	90	90	60	60	90

- Soit la mortalité, soit l'incidence soit les deux sont nulles.  
\* Taux basés sur moins de 10 cas tant pour la mortalité que pour l'incidence

# BIBLIOGRAPHIE

- [1] Waterhouse, J., Muir, C. et al. CANCER INCIDENCE IN FIVE CONTINENTS, Volume II, Centre international de recherche sur le cancer, OMS, Lyon, 1976, p. 3.
- [2] MANUEL DE L'OMS: registres normalisés du cancer (registres hospitaliers), OMS, Genève 1976.
- [3] MANUEL FOR CANCER RECORDS OFFICERS, Institut national du cancer du Canada.
- [4] Op. cit. en [1], pp 45-51.
- [5] Nagurn, D.N., S.G. Currie et B. Heath, ÉVALUATION DE LA QUALITÉ DES STATISTIQUES DE L'ÉTAT CIVIL: étude pilote. Statistique Canada. Division de la santé.
- [6] King, H.S., Wigle, D.T., Hill, G.B., Silins, J., MORTALITY TRENDS FOR CANCERS OF THE CORPUS UTERI AND CERVIX UTERI, Alberta 1969-1978, CMAJ.
- [7] Doll, R., Peto, R., THE CAUSES OF CANCER: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today, JNC1, Vol. 66, N° 6, Juillet 1981.
- [8] Percy, C., COMPARISON OF THE CODING OF DEATH CERTIFICATES RELATED TO CANCER IN SEVEN COUNTRIES, Public Health Reports, Vol. 93, N° 4, Juillet-aout 1978.
- [9] Feigl, P., Breslow, N.E., Laszlo, J., Priore, R.L., Taylor, W.F., U.S. CENTRALIZED CANCER PATIENT DATA SYSTEM FOR UNIFORM COMMUNICATION AMONG CANCER CENTRES, JNC1, Vol. 67, N° 5, Novembre 1981, p. 1019.

fet, les règles actuelles sur la déclaration des cancers primitifs multiples exigent des rapports distincts sur les cancers affectant les deux côtés de la plupart des organes bilatéraux.

Dans l'ensemble, nous croyons que même s'il existe une certaine incohérence provoquée par la déclaration des cancers primitifs multiples et par les déclarations en double, le problème est très négligeable en regard de celui qui est provoqué par la sous-représentation.

#### 4. SOMMAIRE

Les techniques décrites dans cet article ont permis d'identifier les niveaux différents de qualité des données sur l'incidence et la mortalité du cancer. Nous avons trouvé que les ratios mortalité-incidence en particulier peuvent servir à évaluer les erreurs de représentation qui constituent l'une des principales préoccupations des responsables d'un système de bonne qualité pour déterminer l'incidence du cancer. La qualité des données touchant ceux qui sont inscrits dans le système de déclaration des cas de cancer est suffisamment bonne pour qu'il soit possible d'évaluer la qualité de la classification des causes de décès dans le système de mortalité, par le biais d'un appariement de micro-données. En fait, un appariement informatisé de micro-données pourrait servir à évaluer la sous-représentation parce que l'on pourrait déterminer le nombre de décès dus au cancer non inscrits auparavant au Système national de déclaration des cas de cancer.

#### REMERCIEMENTS

Les auteurs tiennent à remercier J. Gorman, D. Lawrence, K. McClean, S. Moore et P. Walsh qui les ont aidés à préparer les données destinées à cet article. Ils sont également reconnaissants à M. J. Silins pour ses commentaires. Ils remercient enfin les membres du comité de lecture pour leurs commentaires.



Les règles sur la déclaration des cancers primitifs multiples sont difficiles à interpréter; aussi doit-on s'attendre à certaines différences à cet égard parmi les provinces.

L'inscription involontaire en double du même cas de cancer peut survenir si un bureau provincial ne peut déterminer si le même cas a été inscrit auparavant (peut-être parce que l'information sur l'identité est insuffisante) ou si le même cas est déclaré par deux bureaux différents<sup>7</sup>. Nous avons limité notre recherche sur les cancers primitifs multiples aux entrées multiples décrivant le même siège du cancer (au niveau à 3 chiffres du code ICDA).

Nous n'avons pas tenté de séparer les inscriptions en double des cas de cancers primitifs multiples, bien que l'on puisse presumer que la majorité des cas déclarés par deux bureaux différents font double emploi alors que les cas déclarés par le même bureau sont davantage susceptibles d'être des cancers primitifs multiples, surtout dans le cas de ceux dont le 4e chiffre du code de diagnostic est différent.

En nous servant de critères très stricts d'appariement et en excluant les cancers de la peau (autres que le mélanome de la peau) nous avons déterminé que 1.7 % (6113) des dossiers formant le fichier sur l'incidence du cancer entre 1969 et 1978 constituaient des inscriptions multiples.

Ce taux variait de 0.5 % à 1.9 % selon la province. Seulement 1.4 % des dossiers multiples avaient été déclarés par deux bureaux différents. Dans 88.4 % des cas, il y avait concordance jusqu'au niveau du 4e chiffre du code de diagnostic.

Quant au siège du cancer, si l'on ne tient compte que des sièges figurant dans plus de cinquante dossiers, le taux des cancers primitifs variait entre 1.0 % pour le cancer de l'estomac et du pancréas et 3.6 % pour le cancer du sein. Le taux élevé en ce qui concerne le cancer du sein n'est pas étonnant; en ef-

<sup>7</sup> Le Système national de déclaration des cas de cancer ne procède pas à des vérifications régulières de ce genre de double emploi.

et 1974-1978), on relève une certaine amélioration au cours de la période plus récente.

Dans les dossiers sur la mortalité du cancer pour la même période, une date de naissance complète existe dans plus de 95 % des cas. Toutefois au moins une partie de ce taux élevé est due au fait que le système de mortalité impute une date de naissance tirée de l'âge et de la date de décès lorsque la date de naissance n'est pas déclarée. En 1976, le taux d'imputation était de 11.5 % (5). On ne procède pas à des imputations de ce genre dans le cadre du système de déclaration des cas de cancer<sup>6</sup>.

Les analyses de petites régions où il se produit des cas de cancer nécessitent une information détaillée sur le lien de résidence. Les données sur la mortalité liée au cancer sont bien plus utiles à ces fins parce que les codes de résidence dans la division de recensement (comté) figurent dans 99.8 % des dossiers et les codes de subdivision de recensement (cité, ville, village) dans 96.2 % des dossiers. Par contre, dans le fichier sur l'incidence du cancer, les codes de division figurent dans 89.6 % des dossiers et les codes de subdivision dans seulement 25.2 % des dossiers. Dans le fichier en question, il y a eu au cours de la deuxième période une amélioration quant à l'information fournie sur la subdivision de recensement.

#### 3.2.4 Inscriptions multiples dans le système de déclaration des cas de cancer (Tableau 7)

La possibilité de comparer les données sur les cas de cancer est affectée s'il existe des différences dans la déclaration des cancers primitifs multiples chez la même personne et dans une inscription involontaire faisant double emploi.

<sup>6</sup> Les imputations peuvent être utiles, à des fins statistiques mais elles sont trompeuses dans les études de post-observation, à moins qu'il ne soit précisé que l'information est basée sur une imputation.

Pour ce qui est des dossiers dont l'appariement a réussi, il y avait concordance du diagnostic d'un cancer primitif du poumon dans 91.4 % des cas, taux très semblable à celui que nous avons trouvé dans la comparaison inverse. Les échantillons destinés aux deux comparaisons avaient été choisis de façon indépendante; il est donc rassurant de constater la concordance des résultats en ce qui concerne le diagnostic. Parmi les cas restants, 4.3 % présentaient la cause sous-jacente du décès attribuée à des sièges de cancer autres que l'appareil respiratoire et 3.4 % une cause qui n'était pas le cancer. Dans le cas de ce dernier groupe, il est possible que le certificat de décès ait mentionné le cancer comme une cause ayant contribué au décès. Cette analyse est fautive, mais nous ne l'avons pas entreprise.

### 3.2.3 Disponibilité et intégralité des données (Tableaux 5 et 6)

Une mesure de la qualité et de l'utilité des fichiers de données est la fréquence d'une information valide pour des données élémentaires spécifiques. Cette mesure est brute parce que le terme "valide", selon la définition que nous vous en donnons ici, signifie valide conformément aux contrôles d'édition informatisés et n'exclut pas l'invalidation de l'information par l'imputation de l'information manquante ou des erreurs de définition ou de classification de la donnée élémentaire.

Sous réserve de ce qui précède, la mesure en question peut se révéler utile en faisant ressortir s'il y a eu des améliorations dans le temps dans la déclaration des données élémentaires et où ces améliorations se sont produites et s'il est possible ou non de soumettre les données à des analyses particulières. Par exemple, l'information sur la date de naissance revêt de l'importance dans les analyses purement statistiques (spécifiques à l'âge) des données ainsi que dans les études post-observation qui dépendent de l'exactitude de l'information sur l'identité des personnes.

Dans le fichier incidence du cancer, une date de naissance complète (jour, mois et année) ne figure en moyenne que dans 68 % des dossiers relatifs à la période 1969-1978. Si l'on considère séparément les deux périodes (1969-1973

des fichiers (ou peut-être le fait qu'un cancer du poumon ait été diagnostiqué pour la première fois avant 1969, suivi par l'inscription ultérieure d'un autre cancer primitif) explique le manque de concordance. Parmi les provinces, le taux de concordance des diagnostics variait entre 74.4% et 100%.

Il est intéressant de noter que pour la province présentant une concordance totale des diagnostics, nous avons aussi enregistré le taux de succès le plus élevé (96.4%) pour repérer les dossiers correspondants dans les deux fichiers. Cela pourrait éventuellement indiquer une liaison étroite entre le bureau provincial de l'état civil et le bureau d'enregistrement des cas de cancer dans la province. Lors d'un appariement inverse d'un échantillon de dossiers sur l'incidence du cancer avec des dossiers sur la mortalité (décrit à la partie 3.2.2), c'est la même province qui présentait le taux le plus élevé d'appariements réussis ainsi qu'une concordance totale en matière de diagnostic.

### 3.3.2 Recherche dans le fichier incidence/mortalité

La recherche inverse, basée sur le fichier incidence pour les années 1969 à 1978, comme point de départ, et visant à repérer un dossier correspondant dans le fichier complet de mortalité pour la même période, a produit un taux de succès de 69.4% de l'échantillon de 504 cas étudiés où le diagnostic indiquait un cancer primitif du poumon. Le taux d'appariements manqués était donc plus élevé que dans le cas des appariements mortalité/incidence pour toutes les provinces et il variait de 19.6% à 44.6%. Les raisons possibles de ce manque d'appariement sont a) soit que le patient était toujours en vie à la fin de l'année 1978, b) soit que l'information sur son identité n'est pas suffisante pour permettre l'appariement. En règle générale, il est moins probable que l'on trouve un dossier correspondant lorsqu'on recherche dans le fichier incidence/mortalité, car certaines personnes chez qui on a diagnostiqué le cancer du poumon survivent effectivement, alors que toutes les personnes qui meurent de ce cancer devraient être inscrites comme des cas nouveaux soit avant le décès soit au moment du décès.



en entreprenant par exemple des recherches sur la précision du codage sur le terrain de la cause du décès et du diagnostic, comme le décrivent deux rapports américains (8) et (9), cela faciliterait l'interprétation des résultats des recherches épidémiologiques.

### 3.2.1 Recherche dans le fichier mortalité/incidence

Pour l'échantillon de 504 enregistrés de décès indiquant le cancer du poumon comme cause sous-jacente de décès entre 1969 et 1978, nous avons trouvé 415 enregistrés correspondants (82 %) dans le fichier incidence du cancer pour les années de diagnostic 1969-1978<sup>5</sup>. Le taux des appariements manqués variait entre 3.6 % et 26.8 % parmi les neuf provinces. Ce taux est influencé par quatre facteurs principaux :

a) le fait que le cancer ait été diagnostiqué pour la première fois avant 1969, b) le fait qu'il reflète une sous-représentation des nouveaux cas, c) le fait que l'information sur l'identité ne suffisait pas pour permettre l'appariement des dossiers, ou d) le fait que le code de la cause du décès indiquait par erreur le cancer. Nous ne disposons pas de données suffisantes pour évaluer le poids relatif de chacun de ces facteurs.

Parmi les 415 enregistrés de décès pour lesquels nous avons trouvé des dossiers correspondants dans le fichier incidence, il y avait concordance du diagnostic (cancer primitif du poumon) dans 92.5 % des cas. Il y avait une concordance de 95.2 % sur la présence d'un cancer de l'appareil respiratoire. Les quelques autres dossiers du fichier "incidence" indiquaient des diagnostics de sièges autres que l'appareil respiratoire. Il est généralement reconnu que l'information sur le diagnostic qui figure dans les registres du cancer est plus précise que l'information sur la cause du décès figurant aux certificats de décès. Toutefois, étant donné l'enversure de cette étude, il n'était pas possible de déterminer si une classification erronée dans l'un ou l'autre

<sup>5</sup> Dans six cas, il existait pour la même personne plus d'un enregistrement correspondant dans le fichier mortalité. Nous avons retenu un seul de ces dossiers comme un appariement réussi.



Les patients plus âgés, la sous-représentation de la population plus âgée est plus marquée dans les bureaux qui disposent généralement de systèmes d'ins-  
cription moins complets.

Les données recueillies au Canada vont donc dans le sens des recommandations de l'Union internationale contre le cancer (1970) et du Centre international de recherche sur le cancer (1976) et de la réitération de la recommandation faite dans un récent article de Doll et Peto (7), à savoir "que l'on n'obtient des comparaisons suffisamment dignes de foi (de l'incidence du cancer) que si ces comparaisons se limitent aux hommes et aux femmes de l'âge moyen".

### 3.2 Appariement des dossiers mortalité et incidence (Tableaux 3 et 4)

Comme Statistique Canada est responsable de la gestion des fichiers de données tant sur l'incidence que sur la mortalité attribuable au cancer, il est possible de comparer les inscriptions sur les personnes qui figurent dans les deux registres distincts pour vérifier l'information qui y est consignée.

Pour cette partie de l'étude sur la qualité des données, la précision de la détermination du diagnostic et de la cause du décès présentait un intérêt particulier. Dans le cadre de l'étude, on ne peut que décrire les résultats - les raisons qui expliquent les divergences découvertes demeurent inconnues. Toutefois, nous estimons que les résultats sont révélateurs et qu'ils indiquent que, dans le cas du cancer particulier choisi aux fins d'analyse (cancer du poumon), la concordance de diagnostics entre les deux fichiers est généralement élevée (plus de 90 %).

Notre étude indique aussi qu'il serait possible de procéder à un appariement sur une plus grande échelle pour évaluer la précision des codes de diagnostic et de cause de décès. Évidemment, si une étude sur une plus grande échelle était basée sur un échantillon, il serait préférable de stratifier celui-ci selon l'année de diagnostic ou l'année de décès. En théorie, si l'on se sert de techniques informatisées d'appariement, il est possible d'appliquer ce type d'analyse à tous les sièges de cancer. Si l'on voulait compléter cette étude

Parmi les principaux sièges du cancer qui ont été examinés pour les hommes, les cancers du poumon et de l'estomac allaient de pair avec les ratios mortalité-incidence les plus élevés dans tous les bureaux, mais la variation interprovinciale des ratios était la plus élevée pour les cancers du côlon (rectum exclu), de la prostate et de la vessie. Dans les études visant à mettre en relief les différences de risque du cancer selon la région géographique (province), l'emploi des données sur l'incidence du cancer constituerait donc une méthode plus fiable dans le cas des deux premiers sièges de cancer cités.

Pour les femmes, parmi les principaux sièges examinés, les cancers du côlon et de l'ovaire correspondaient aux ratios les plus élevés dans tous les bureaux. La variation interprovinciale des ratios était la plus élevée pour le cancer de l'utérus et du col de l'utérus ainsi que pour le cancer du côlon. Aux fins de comparaisons interprovinciales, les données sur l'incidence du cancer du sein et du cancer de l'ovaire seraient donc plus fiables.

Dans le cas des sièges du cancer de l'utérus (autre que celui du col) et du cancer du col, les ratios indiquent de grandes variations interprovinciales si l'on tient compte des sièges séparément. La variation se réduit considérablement si l'on combine les deux sièges, ce qui laisse entendre qu'il existe des différences dans la précision du diagnostic et la détermination de la cause du décès dans le cas de ces deux sièges.

Pour les principaux groupes d'âge, nous avons examiné les ratios mortalité-incidence spécifiques au siège. Les taux les plus élevés se sont invariablement manifestés dans le cas des personnes âgées (65 ans et plus) dans tous les bureaux et pour tous les sièges indiqués. Il s'agit là d'un phénomène normal puisque le risque de décès augmente avec l'âge; proportionnellement, il y a donc plus de décès que de nouveaux cas de cancer parmi les personnes âgées. Il est aussi reconnu qu'il est généralement plus difficile de diagnostiquer et de déclarer les cancers chez les personnes âgées. Toutefois, l'augmentation relative des ratios dans le cas de ces personnes est bien plus élevée pour les bureaux qui ont, au départ, les ratios moyens les plus élevés. Cela indique que même si tous les bureaux peuvent éprouver certaines difficultés à inscrire

général, les bureaux qui se servent d'un plus grand nombre de sources différentes d'inscriptions n'ont pas plus d'inscriptions multiples pour le même siège que les bureaux qui se servent d'un nombre moins élevé de sources.

Les ratios mortalité-incidence du cancer dans les six autres bureaux se ressemblaient davantage. Dans ces bureaux, nous n'avons pas relevé une tendance systématique selon laquelle un bureau présentait des rapports plus élevés ou moins élevés pour tous les sièges et pendant toutes les deux périodes à l'étu-

de.

Il existe de nombreux facteurs qui peuvent influencer les variations des ratios observés selon le site du cancer. Parmi les facteurs qui tendent à produire une inscription moins complète des nombreux cas - et par conséquent, des ratios mortalité-incidence plus élevés - il faut citer la difficulté de diagnostiquer le cancer (par exemple dans les organes profondément enfouis) et le manque d'accès à des sources spécifiques de données (tels que les rapports d'hématologie confirmant un diagnostic de leucémie).

Les facteurs qui peuvent provoquer la surinscription des nouveaux cas et des rapports moins élevés sont les programmes de dépistage de masse (qui peuvent aboutir à l'inclusion des cas répandus, surtout dans le cas des tumeurs à croissance lente), les inscriptions qui se chevauchent, l'inclusion de cancers in situ ainsi que l'inclusion de cancers latents découverts seulement au moment de l'autopsie (tel est notamment le cas du cancer de la prostate). Par ailleurs, les différences de précision dans la détermination du diagnostic ou de la cause du décès peuvent mener à des différences artificielles. Par exemple, les certificats de décès peuvent porter l'inscription "cancer de l'utérus, non spécifié" ou "leucémie, non spécifié" comme étant la cause du décès alors qu'un bureau d'inscription du cancer disposera souvent d'une information plus précise et affectera des codes plus précis (6). C'est pourquoi une analyse des ratios mortalité-incidence au niveau du diagnostic plus détaillé pourrait révéler des divergences importantes.



siège, un sexe ou un groupe d'âge donné, il faut se douter qu'il y a des différences sur le plan de l'intégralité des inscriptions de nouveaux cas de cancer. Un rapport plus élevé, qui signifie une proportion plus élevée de décès en regard des nouveaux cas pendant la même période peut indiquer une insc-

tion moins élevée de ces cas.

Pendant toutes les deux périodes, il y avait deux bureaux où l'on relevait constamment les taux les plus élevés pour tous les sièges réunis et pour la plupart des principaux sièges indiqués. Il n'y a guère de doute que ces taux élevés sont dus à une sous-représentation de nouveaux cas; en effet, ces bureaux sont les seuls qui ne se servent pas de certificats de décès comme sources d'inscription. En outre, l'un de ces bureaux ne se sert que d'une seule source de données, à savoir les rapports des hôpitaux, pour inscrire les cas de cancer. Le bureau en question avait auparavant présenté les conclusions d'une étude spéciale qui montrait que ce bureau ne recevait des avis que sur un pourcentage estimé à 70 % des victimes du cancer admises dans les hôpitaux jusqu'à la fin de 1976. À la suite de cette étude, d'importants changements ont été apportés en vue d'améliorer le système de déclaration. Depuis 1977, ce bureau déclare un nombre plus élevé de cas de cancer, ce qui se traduit par une réduction marquée des ratios mortalité-incidence.

Tous les autres bureaux d'inscription du cancer au Canada se servent des sources multiples de déclaration, ce qui est considéré essentiel pour couvrir un vaste champ d'observation et ce qui pourrait avoir une incidence positive sur l'intégrité et la qualité des données élémentaires. Pour mener notre étude, nous avons examiné l'intégrité des données élémentaires déclarées (voir sous-section 2.3). Toutefois, il se trouve que le seul bureau qui a recours à une seule source de données occupe en fait un rang passablement élevé sur le plan de l'intégrité de l'information touchant de nombreuses données élémentaires.

Le recours à des sources multiples d'inscription comporte un risque: le double emploi. Toutefois, l'analyse des inscriptions multiples concernant la même personne et le même siège de cancer ne confirme pas ce risque. En

Parmi ces dossiers, nous avons relevé les inscriptions multiples comme suit:

- L'année, le mois et le jour de naissance étaient indiqués et concord-  
daient, ou
- le jour de naissance n'était pas indiqué sur au moins un dossier mais  
les mois de naissance concordent.

Nous avons aussi vérifié à la main tous les groupes comprenant au moins trois  
personnes où l'information sur le mois et le jour ne concordait pas et tous  
les groupes de deux personnes où l'information sur le mois ou le jour manquait  
dans au moins un dossier.

En tout, nous avons fini par relever 6,113 dossiers susceptibles de faire dou-  
ble emploi. Ces dossiers correspondent à 5,947 personnes. (À noter que cer-  
taines personnes faisaient double emploi plus d'une fois). Nous n'avons pas  
jugé s'il s'agissait là d'inscriptions multiples valables ou effectivement de  
doubles emplois.

Pour chaque valeur ICDA-8 de 3 chiffres, nous présentons dans le tableau 7 la  
ventilation de ces doubles emplois possibles selon qu'il s'agit de dossiers  
provenant du même bureau ou de bureaux différents et s'il s'agit du nombre de  
doubles emplois possibles qui ont le même quatrième chiffre de la classifica-  
tion ICDA-8.

### 3. DISCUSSION

#### 3.1 Ratios mortalité-incidence (tableaux 1 et 2)

Les ratios mortalité-incidence du cancer peuvent fournir une indication sur  
l'intégralité des inscriptions. Dans tous les bureaux, ces taux varient selon  
le siège du cancer (les taux les plus élevés correspondant aux taux de survie  
les moins élevés), et selon l'âge et le sexe. Toutefois si une comparaison  
des ratios des différents bureaux indique de grandes différences pour un



provinces qui ont la déviation la plus élevée par rapport à la moyenne nationale. Pour le dossier mortalité, nous indiquons les résultats uniquement pour les décès dus au cancer dans les neuf provinces autres que l'Ontario afin que la comparaison avec le SNDC soit plus valable.

## 2.4 Inscriptions multiples au système de déclaration des cas de cancer

Le principe du Système national de déclaration des cas de cancer est de consigner tous les cas nouveaux de néoplasmes malins. Une personne devrait être inscrite plus d'une fois lorsqu'il se développe chez elle des néoplasmes malins multiples. Afin d'éviter la double inscription du même cas ou la double déclaration des patients inscrits dans plus d'une province, tous les responsables des registres provinciaux de cas de cancer suivent des procédures réglementaires. Il peut arriver malgré cela que le même cas de cancer soit déclaré deux fois. Dans le but d'évaluer l'ampleur des doubles emplois, nous avons recherché les dossiers susceptibles de faire double emploi. Notre recherche a été loin d'être exhaustive, de sorte que le nombre de doubles possibles que nous avons trouvés est en deçà du chiffre réel. Parmi les 457,158 dossiers, nous avons éliminé ceux qui indiquaient des noms ou des années de diagnostic invalides. Pour les dossiers où il manquait l'année de naissance, nous avons déterminé l'année en question à partir de l'âge si celui-ci était connu. Nous avons aussi éliminé les dossiers sur le cancer de la peau (ICDA-8 : 173) puisque l'on sait que ce type de cancer apparaît à de multiples occasions.

Parmi les dossiers restants, nous avons retenu ceux qui remplissaient toutes les conditions suivantes :

- Concorde parfaite des années de naissance ou des années de naissance calculées
- concordance parfaite des noms
- exactitude des quatre premières lettres du prénom
- conformité du code de trois chiffres ICDA-8.

Dans les tableaux 5 et 6, nous présentons les moyennes nationales pour les deux fichiers centraux; nous indiquons également les valeurs pour les

tité a changé de manière appréciable au cours des années ultérieures.

bles de cinq ans (1969-1973 et 1974-1978) afin de pouvoir observer si la qualification de variation. Nous avons groupé les fréquences relatives en deux ensembles dans une variable numérique ou que la valeur numérique dépasse l'intervalle des données invalides apparaissent lorsque l'on trouve des caractères alphabétiques en données valides et en données invalides. Outre les valeurs nulles, petites régions. Pour chaque élément d'information, nous classifions les données de 2.2) ou de créer des tabulations spéciales telles que des statistiques sur de difficile d'appartier des dossiers provenant d'autres fichiers (ex. partie Nous choisissons ces éléments pour souligner à quel point il serait facile ou

- date de naissance (jour, mois, année)
- âge
- lieu de naissance
- comté et subdivision de résidence

sur les éléments d'information suivants:

La fréquence relative de données valides pour des éléments d'information spécifiques offre une mesure simple de la qualité des fichiers de données. Dans le cas du système national de déclaration des cas de cancer et dans le cas des décès dus au cancer consignés dans le fichier mortalité, nous nous concentrons

## 2.3 Disponibilité et intégralité des données

Les taux moyen d'appariements réussis a été de 69.4 %. Les taux variaient entre 55.4 % et 80.4 % parmi les neuf provinces. Sur les appariements réussis, 92.0 % avaient la même classification à 4 chiffres (ICDA-8). Ces taux variaient entre 74.4 % et 100.0 %. Pour les provinces ayant les taux d'appariements les plus élevés ou les moins élevés comportant la même classification ICDA-8, nous indiquons au tableau 4 la ventilation des causes observées des décès classifiés.

## 2.2.2 Rapport incidence-mortalité

Nous avons aussi choisi un échantillon de dossiers tirés du SNDCC et les avons apparés au fichier mortalité. Nous avons choisi au hasard dans chacun des bureaux de l'état civil 56 dossiers indiquant des néoplasmes malins du poumon et des bronches (ICDA-8: 162.1) et rassemble ainsi un échantillonnage de 504 dossiers. Nous avons ensuite vérifié les dossiers apparés pour déterminer la cause sous-jacente du décès d'après le fichier complet mortalité de 1969 à 1978. Pour cette étude, nous n'avons pas vérifié la cause du décès figurant sur le certificat original de décès, mais on pourrait procéder à cette vérification pour les prochaines études.

Les résultats tirés de cet appariement manuel peuvent être classés comme suit:

- (a) on ne trouve pas de dossier apparé,
- (b) on trouve un appariement, mais la cause du décès est autre que le cancer,
- (c) on trouve un appariement où le cancer est la cause du décès mais non le cancer du poumon ou des bronches,
- (d) on trouve un appariement indiquant la même cause de décès.

Si l'on ne trouve pas d'appariement, cela indique l'insuffisance du processus d'appariement, à moins que le décès n'ait eu lieu après 1978 ou à l'extérieur du Canada ou encore que la personne soit toujours en vie.

Si l'on trouve un appariement où la cause sous-jacente du décès est différente, cela indique l'une des possibilités suivantes:

- (a) un risque concurrent a prévalu,
- (b) le cancer a été une cause contribuant au décès, mais non la cause sous-jacente, ou encore
- (c) la cause sous-jacente du décès indiquée par le fichier central mortali-

té était inexacte, ou bien le siège du cancer indiqué par le SNDCC était inexact (ce dernier cas étant censé être moins probable).

- (a) on ne trouve aucun dossier apparié
- (b) on trouve un appariement mais le dossier indique un siège différent pour le cancer
- (c) on trouve un appariement et le dossier indique le même siège pour le cancer.

Si on ne trouve aucun appariement cela indique qu'il y a une sous-représentation, ou que le cancer a été diagnostiqué pour la première fois en Ontario ou à l'extérieur du Canada ou bien avant 1969, ou encore que le décès n'était pas réellement attribuable au cancer. Il est possible aussi comme on l'a déjà indiqué, que le processus d'appariement proprement dit ne soit pas parfait. Si l'on trouve un appariement, mais que les dossiers indiquent des sièges différents pour le cancer, cela peut indiquer une erreur soit dans le fichier mortalité, soit dans le SNDCC. Comme nous l'avons déjà mentionné, on croit généralement que le dossier mortalité offre une classification plus précise des maladies en raison du taux élevé de confirmations histologiques.

Nous avons entrepris une étude à échelle réduite pour mesurer ce phénomène. Dans chacune des provinces, à l'exception de l'Ontario, nous avons choisi un échantillon aléatoire de 56 dossiers indiquant un néoplasme malin du poumon ou des bronches comme la cause sous-jacente du décès (ICDA-8: 162.1) et rassemble ainsi un échantillonage de 504 dossiers. Nous avons retenu seulement les décès survenus entre 1969 et 1978.

À l'échelle nationale, le taux d'appariements réussis était de 82.3 %. Les taux variaient entre 73.2 % et 96.4 % parmi les neuf provinces. Sur les appariements réussis, 92.5 % avaient la même classification à 4 chiffres (ICDA-8). Ces taux variaient entre 74.4 % et 100 %. Pour les provinces ayant les taux d'appariements les plus élevés ou les moins élevés comportant la même classification ICDA-8, nous indiquons au tableau 3 la ventilation des classifications de maladies observées.



Comme points de départ, nous avons choisi dans les deux fichiers les dossiers faisant état de néoplasmes malins du poumon ou des bronches (ICDA-8<sup>4</sup> : 162.1) pour les années 1969 à 1978. Ce choix a été basé sur des considérations d'incidence élevée, de mortalité élevée et de courtes périodes de survie; les conditions étaient donc favorables pour trouver des dossiers apparus dans les deux fichiers. Malgré cela, en raison du temps écoulé entre le diagnostic et décès il est vrai que les décès provoqués par le cancer au cours des premières années et les cancers nouvellement diagnostiqués au cours des années ultérieures sont moins susceptibles de faire l'objet d'un dossier correspondant dans le fichier apparié. L'analyse des résultats serait améliorée dans les prochaines études si la conception de l'échantillon permettait de contrôler l'année du décès et l'année du diagnostic.

Nous avons choisi deux échantillons indépendants: l'un du fichier de la mortalité et l'autre du fichier du Système national de déclaration des cas de cancer (SNDCC).

## 2.2.1 Rapport mortalité-incidence

Tous les décès dus au cancer entre 1969 et 1978 devraient avoir, du moins en théorie, un dossier correspondant dans le fichier du SNDCC. Les principales exceptions à cette règle sont les cas où le cancer a été diagnostiqué pour la première fois en Ontario ou à l'extérieur du Canada ou bien les cas où le cancer a été diagnostiqué pour la première fois avant 1969. Observation faite de ces exceptions, l'absence d'un dossier correspondant dans le SNDCC indique une sous-représentation. On suppose, bien entendu, que la cause sous-jacente du décès qui figure dans le fichier mortalité est dépourvue d'erreur.

Par conséquent, si l'on choisit un échantillon de décès dus au cancer et si on les apparie au SNDCC, on est en présence d'un certain nombre de possibilités:

<sup>4</sup> International Classification of Diseases (Classification internationale des maladies) adaptée pour les États-Unis, huitième révision.



Si les ratios mortalité-incidence sont trop différents pour la plupart des sièges, on pourra négliger de tenir compte des différences de taux en a) et b). Quant aux taux d'erreur (c), les études sur l'erreur de codage de la cause sous-jacente du décès ont fait ressortir des taux d'erreur de moins de 10 % (5). Dans un rapport non publié, on a trouvé que ces taux d'erreur pourraient varier entre 3 % et 18 % parmi les bureaux de l'état civil. Toutefois, les différences observées parmi les ratios mortalité-incidence (voir tableaux 1 et 2) ne peuvent entièrement s'expliquer par ces taux d'erreur. Par ailleurs, comme presque 90 % des inscriptions de nouveaux cas de cancer sont confirmées sur le plan histologique, le taux d'erreur en d) est négligeable. Aussi, ces ratios mortalité-incidence donnent-ils une indication sur l'erreur de représentation dans le Système national de déclaration des cas de cancer.

En nous concentrant sur les sièges les plus fréquemment diagnostiqués (à l'exception du cancer de la peau) d'après le fichier SNDCC, nous présentons aux tableaux 1 et 2, pour chaque sexe, les ratios mortalité-incidence à l'échelle nationale ainsi que les provinces comptant les ratios les plus élevés ou les moins élevés, pour deux périodes de cinq ans, ainsi qu'une ventilation par groupes d'âge. Nous excluons l'Île-du-Prince-Édouard de notre étude parce que le nombre de cas observés dans cette province est trop faible pour permettre une comparaison valable.

## 2.2 Appariement des dossiers mortalité et incidence

Pour établir la possibilité d'évaluer les erreurs de classification des causes de décès, ou les erreurs de classification des sièges du cancer, on peut tirer un échantillon des inscriptions provenant soit du fichier mortalité, soit du fichier incidence, et ensuite rechercher dans l'autre fichier des dossiers appariés. La recherche manuelle ne garantit pas que l'on trouvera tous les appariements valables et, en fait, le taux d'appariement peut être différent d'un bureau de l'état civil à l'autre, parce que le niveau de détail des variables d'appariement peut varier d'un bureau à l'autre. (Voir à la partie 2.3 une étude sur la disponibilité des données).

## 2. DESCRIPTION DES MESURES

Dans cette partie, nous décrivons certaines méthodes pour étudier la qualité des données contenues dans les registres sur la mortalité et l'incidence du cancer.

### 2.1 Ratios mortalité-incidence

En vue d'étudier les taux relatifs de sous-représentation parmi les inscriptions de l'incidence du cancer, nous considérons les ratios entre les décès et les cas de cancer. Étant donné qu'au Canada tous les décès sont enregistrés selon la cause du décès et que suivant le "système national de déclaration des cas de cancer" (SNDCC), tous les nouveaux cas de néoplasme malin sont enregistrés, si les deux systèmes de consignation étaient de la même qualité dans tous les bureaux de l'état civil on devrait s'attendre à ce que la proportion entre les taux de mortalité et les taux d'incidence relatifs à un siège donné soit passablement uniforme au sein d'une population d'un âge et d'un sexe données et au cours de périodes suffisamment longues. (Nous calculons ces proportions de décès et de cas pour des périodes de cinq ans et de dix ans afin de réduire l'effet du décalage dans le temps entre la déclaration d'un cas de cancer et le décès). Cette proportion manquerait de cohérence si l'un ou l'autre des taux ci-dessous différerait d'un bureau de l'état civil à l'autre.

- (a) taux de survie ou accroissement soudain du taux d'incidence,
- (b) taux de mortalité associée à d'autres risques concurrents,
- (c) taux d'erreur de codage de la cause sous-jacente de décès
- (d) taux d'erreur dans la classification du siège du cancer dans le cas des nouveaux cancers
- (e) taux de sous-représentation ou de sur-représentation des cas de cancer

<sup>3</sup> Exceptionnellement, on enregistre les cancers métastatiques lorsque le siège primaire est inconnu.

du cancer et le sous-ensemble de dossiers tirés du fichier sur la mortalité, lequel fait apparaître le cancer comme cause sous-jacente du décès.

Voici les aspects de la qualité des données que nous avons retenus pour notre étude :

1. L'intégralité des inscriptions de nouveaux cas de cancer par le biais d'une comparaison entre la mortalité attribuable au cancer et l'incidence du cancer pendant la même période. Cette comparaison constitue un indicateur très approximatif (mais couramment utilisé) de l'intégralité des inscriptions.

2. La compatibilité des codes de détermination du diagnostic et de la cause du décès par le biais d'un appariement des dossiers individuels formant les deux fichiers.

3. La disponibilité et l'intégralité des données élémentaires par le biais d'une analyse de la fréquence à laquelle des valeurs valides apparaissent dans les fichiers.

4. L'inscription des cas de cancer primitifs multiples par le biais d'un appariement des dossiers formant le registre.

Notre étude couvre la période comprise entre 1969 et 1978 pour laquelle on dispose des données sur l'incidence du cancer. Nous avons exclu l'Ontario du champ de toutes nos recherches étant donné que le Système national de déclaration des cas de cancer comprend des données sur cette province uniquement pour la période qui s'étend de 1969 à 1971<sup>2</sup>.

La 2<sup>e</sup> partie de cet article expose la méthode adoptée pour les recherches et la 3<sup>e</sup> partie les conclusions de l'étude.

<sup>2</sup> L'Ontario a créé un système passif d'inscription qui fait appel aux rapports sur les victimes du cancer, préparés à d'autres fins. Cette province prépare actuellement les données relatives aux dernières années.

d'autres aspects de la qualité des données telles que le détail de l'information sociodémographique et géographique qui est fournie. Par ailleurs, les données sur l'incidence du cancer sont sensibles à des facteurs tels que les programmes de dépistage de la masse qui mènent à l'inclusion de cas répandus non diagnostiqués auparavant.

Les rédacteurs de "Cancer Incidence in Five Continents" utilisent et expliquent plusieurs indices qui peuvent servir à déterminer si les données corrigées sont complètes et fiables (4). Parmi ces indices, il faut citer les ratios mortalité-incidence du cancer en tant qu'indicateur de l'intégralité des inscriptions (voir parties 2.1 et 3.1 de cet article).

La déclaration des décès est une obligation légale dans la plupart des pays développés; on suppose donc que les erreurs d'enregistrement y sont minimales. Aux fins de la recherche épidémiologique sur le cancer, les limitations connues ou présumées des données sur la mortalité comprennent un manque d'information sur les cancers non mortels, une information moins précise sur le diagnostic et des erreurs de classification plus fréquentes attribuables à l'explication et au codage de la cause sous-jacente des décès, d'où une précision moindre en comparaison de l'information sur le diagnostic dans les statistiques sur l'incidence du cancer.

Une évaluation de la qualité des statistiques de l'état civil entrepise sous forme d'étude pilote fournit certaines indications sur la qualité des données concernant la mortalité au Canada (toutes les causes), notamment sur les taux d'erreur dans le codage de la cause sous-jacente du décès. Ce taux d'erreur s'élevait à 7.2 % en 1976. Près des deux tiers des erreurs portaient sur le premier et le deuxième chiffres du code de la cause du décès, lequel comprend 4 chiffres. La variation du taux d'erreur selon les causes spécifiques du décès n'a pas été étudiée.

Dans le présent article, nous évaluerons la possibilité de mesurer certains aspects touchant la qualité des deux fichiers de données de Statistique Canada qui servent aux études épidémiologiques, à savoir le fichier sur l'incidence



2. Données sur l'incidence du cancer à l'échelle nationale basées sur les déclarations des registres provinciaux des cas de cancer. Cette série de données a été établie en 1921.

La bonne qualité des statistiques fournissant des renseignements fiables sur les différentielles de risque dépend de l'intégrité et de la précision des cas de cancer enregistrés et de la comparabilité des données entre les différents bureaux de l'état civil et les différentes périodes. En tant qu'indice du risque réel lié au cancer, chacune des données sur l'incidence du cancer et la mortalité due à cette maladie présente de l'intérêt tout en ayant des limites qui lui sont propres.

Les données sur l'incidence du cancer se prêtent particulièrement bien à l'étude épidémiologique de cette maladie parce qu'elles fournissent des renseignements sur tous les cancers et non uniquement sur ceux qui sont mortels, parce qu'elles permettent de prévoir les problèmes naissants et parce que l'information sur le diagnostic est généralement détaillée et de très bonne qualité. Par exemple, la publication "Cancer Incidence in Five Continents" (1) "souligne le rôle que jouent les comparaisons de cette incidence à l'échelle internationale en fournissant des indices sur les causes du cancer en dépit de certaines limitations bien connues des données. Parmi ces restrictions, citons la difficulté d'obtenir la déclaration de tous les nouveaux cas de cancer et les différences entre les bureaux de l'état civil lorsque les nouveaux cas sont bien déclarés. Les principaux facteurs qui influent sur le relevé effectif des cas sont les suivants: quels sont le nombre et les types de sources de données utilisées, dans quelle mesure on procède à une recherche active des cas, depuis combien de temps le bureau de l'état civil existe et si la déclaration des cas de cancer constitue ou non une obligation légale dans la région où est situé le bureau. Au Canada, les bureaux d'enregistrement des cas de cancer emploient une méthodologie passablement hétérogène pour la collecte de leurs données, mais ils essaient tous de suivre les directives tant internationales (2) que nationales (3) pour la consignation normalisée des données sur l'incidence du cancer. Les différences de sources et de techniques influencent non seulement le champ d'observation mais aussi



CERTAINS ASPECTS DE LA QUALITÉ DES STATISTIQUES SUR LA  
MORTALITÉ DUE AU CANCER ET L'INCIDENCE DE CETTE MALADIE

D. Binder et A. Malhotra<sup>1</sup>

Depuis 1921, Statistique Canada, organisme central de regroupement de statistiques au Canada, compile des données sur la mortalité à l'échelle nationale, notamment celles qui touchent la mortalité due au cancer. Il dispose également de données sur l'incidence du cancer qui remontent à 1969.

On peut évaluer de diverses façons la qualité des données de ces fichiers. Les rapports entre la mortalité due au cancer et l'incidence de cette maladie donnent certaines indications sur les erreurs de représentation. L'appariement des micro-données entre les fichiers "incidence" et "mortalité" donnent un aperçu des erreurs de classification. De même, les inscriptions multiples de l'incidence du cancer posent le problème du double emploi. Par ailleurs, l'intégralité et la disponibilité des données élémentaires revêtent de l'importance dans le cas d'études spéciales.

Dans cet article, nous étudions la possibilité de nous servir de ces mesures de la qualité des données et les conséquences que peuvent avoir les mesures en question.

1. INTRODUCTION

Les statistiques démographiques du cancer servent de base aux recherches épidémiologiques sur la distribution et les déterminants de cette maladie ainsi qu'aux programmes de santé visant la prévention, le diagnostic et le traitement du cancer. Statistique Canada, organisme central de regroupement de statistiques, compile deux types de statistiques de ce genre sur le cancer: 1. Données sur la mortalité à l'échelle nationale basées sur les rapports provenant des registres provinciaux de l'état civil. Ces données remontent à 1921.

<sup>1</sup> D. Binder, Division des méthodes d'enquête-institutions et agriculture, Statistique Canada et A. Malhotra, Division de la santé, Statistique Canada.

## BIBLIOGRAPHIE

- [1] Burchell, J.M.: "International Air Passenger Origin/Destination Projection"; exposé présenté à la Division des transports et des communications de Statistique Canada, mars 1981.
- [2] "International Air Passenger O & D Statistics System - User Requirements Specifications"; document du Centre des statistiques de l'aviation (CSA) de la Division des transports et des communications de Statistique Canada, revu en mars 1981.
- [3] "International Air Passenger O & D Statistics System - Feasibility Study Report"; document du CSA, janvier 1981.
- [4] "International Air Passenger O & D Statistics System - Requirements for Functional Specifications (Concepts)"; document du CSA, revu en septembre 1981.
- [5] Rosen, F.G. and Conroy, P.F.: "A Test of the Assignment Technique Based on a Survey of International Air Travelers at Montreal and Toronto"; rapport confidentiel établi pour la Direction générale des études économiques du transport de passagers et de l'aviation, Direction de la recherche, Commission canadienne des transports, février 1977.
- [6] Carpenter, R.: "Specifications for Estimates of the Reliability for the Air Scheduled International Passenger Origin and Destination Estimates"; document de la Division des méthodes d'enquêtes-entreprises de Statistique Canada, revu en janvier 1982.

d'échantillonnage est inférieure à 10% quand il est nécessaire d'inclure des données imputées dans l'estimation du nombre global de passagers dans un marché. Ainsi, l'inclusion de  $f_j$  au lieu de 0.10 dans l'expression de la variance d'une estimation produit un coefficient ajusté qui s'applique aux estimations globales relatives à un marché.

Cette méthode tient compte du fait que des données d'échantillonnage sont utilisées dans la technique d'attribution, mais elle est fondée, comme le procédé qui détermine les estimations elles-mêmes, sur l'hypothèse que les voyages DOD tronqués représentent bien le trafic sans correspondance.

## 8. PROJETS À L'ÉTUDE

Le besoin de disposer de renseignements sur les marchés des transports aériens pour les négociations bilatérales sur les services internationaux a été expliqué dans une des sections précédentes. L'échange de statistiques constitue un des éléments négociables prévus dans les accords sur le transport aérien, et des ententes permettant de tels échanges existent déjà entre le Canada et plusieurs pays. Les notions que ces pays appliquent dans leurs enquêtes ainsi que la qualité de leurs données nous laissent croire que ces chiffres pourraient être utilisés dans le système d'estimation de l'ASIPOD. De telles données fourniraient un échantillonnage des trajets non dénombrés dans l'univers des billets émis pour les vols internationaux. Des études de faisabilité actuellement en cours tentent de déterminer si le nombre de pays qui pourraient échanger ces statistiques contribuerait à améliorer la précision d'un nombre d'estimations suffisant pour justifier l'expansion du système d'estimation de l'ASIPOD en vue d'inclure des "données échangées".

## REMERCIEMENTS

Les auteurs remercient leurs collègues de la Division des transports et des communications et de la Division des méthodes d'enquêtes-entreprises de Statistique Canada pour les nombreuses remarques pertinentes faites pendant la rédaction de ce document.

## 7.2 Variance des estimations globales relatives à un marché

La méthode de calcul des coefficients de variation pour l'ensemble d'un marché doit assurer que les données imputées sont une fonction des données de l'échantillon. Autrement dit, des échantillons différents produisent des valeurs différentes des données imputées et, en conséquence, des grandeurs différentes des estimations globales relatives à un marché. La réutilisation de certaines parties de l'échantillon a un effet sur la distribution d'échantillonnage des estimations.

La méthode que nous proposons corrige la fraction d'échantillonnage de 10% pour qu'elle corresponde au pourcentage de passagers réellement échantillonnés dans chaque domaine d'intérêt. Comme on prend pour acquis l'application de la technique d'attribution, on suppose également que les voyages DOD tronqués représentent bien le trafic sans correspondance. Notre mesure de fiabilité constitue une mesure de précision seulement si cette hypothèse est valable. Il est donc raisonnable de rectifier les poids des voyages DOD échantillonnés afin de prendre en considération le trafic sans correspondance. Ainsi, la fraction d'échantillonnage nécessaire pour l'estimation du  $j^e$  domaine d'intérêt serait  $f_j$ , comme définie dans l'équation (1) plus haut. Dans l'équation (2),  $f_j$  remplacerait  $f$  pour le calcul de  $\hat{\text{var}}(d)$ , ce qui donne la formule pour la variance des estimations globales relatives à un marché :

$$\hat{\text{var}}(\hat{d}_j^1) = \frac{f_j^2}{(1-f_j)} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}.$$

Ensuite, le coefficient de variation des estimations globales relatives à un marché serait :

$$\text{cv}(\hat{d}_j^1) = \sqrt{\hat{\text{var}}(\hat{d}_j^1)} / \hat{d}_j^1.$$

Signalons que  $f_j < 0.10$  lorsque  $a_j > 0$ . Ce qui signifie que la fraction

L'enquête sur l'origine et la destination des passagers payants utilise effectivement un échantillon aléatoire simple de 10%, puisque

(i) le choix des volets dont le numéro d'ordre se termine par "0" produit un échantillonnage systématique, et

(ii) la distribution des billets ne comporte aucun cycle qui causerait une relation entre le nombre de passagers estimé par l'enquête et le dernier chiffre du numéro d'ordre.

L'estimation de la variance peut donc s'écrire :

$$\hat{\text{var}}(d) = N^2 \frac{1}{n} (1-f) v_s$$

$$\text{ou } v_s = \frac{1}{n} \sum_{i=1}^{n-1} (x_i - \bar{x})^2$$

$$\text{et } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

et le coefficient de variation peut s'exprimer sous la forme :

$$cv(\hat{d}) = \sqrt{\hat{\text{var}}(d)/\hat{d}}.$$

A noter que  $\hat{\text{var}}(d)$  peut s'écrire également :

$$\hat{\text{var}}(d) = \frac{f^2}{(1-f)} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

où n est supposé suffisamment grand pour que  $n/(n-1)$  égale approximativement 1.



Prenons les estimations du nombre total de passagers sortants et entrants, sans distinguer la direction du voyage. Parmi les diverses paires d'origines et de destinations, le billet en question ajouterait des passagers aux domaines suivants:

Domaine d'intérêt	Nombre de passagers
Winnipeg - Londres	1
Toronto - Londres	1
Canada - Londres	2
Canada - Europe	2
Est du Canada - Europe	1
Ouest du Canada - Europe	1

Notons que le nombre de passagers par billet dépend du niveau de répartition géographique et, par conséquent, de l'origine et de la destination (c'est-à-dire des domaines d'intérêt) pour lesquelles des estimations du trafic sont calculées.

On peut trouver  $\hat{d}_i$ , l'estimation du nombre de passagers qui voyagent entre une origine et une destination en particulier, d'après l'enquête sur l'origine et la destination des passagers payants, à l'aide de la formule suivante:

$$\hat{d}_i = \frac{\sum_{j=1}^n f_{ij}}{x_i}$$

où  $x_i$  est le nombre de voyages simples appartenant à une paire d'origines et de destinations pour le  $i^e$  billet.

$n$  est le nombre total de billets dans l'échantillon, et

$N$  est le nombre total de billets dans la population de l'enquête sur l'origine et la destination des passagers payants.

$f$  est la fraction d'échantillonnage, c'est-à-dire,  $n/N = 0.1$ .

## 7. ESTIMATION DE LA VARIANCE

L'application de la technique d'attribution à l'estimation des marchés internationaux représente une solution nouvelle pour un problème de taille. Elle ne règle pas complètement les difficultés causées par la non-exhaustivité du dénombrement, mais elle constitue un grand pas en avant, comme d'ailleurs la production de statistiques sur la variance des estimations. Les chiffres sur la variance devraient à la fois tenir compte de la technique d'attribution et de ses hypothèses, et fournir une mesure claire de la fiabilité des estimations relatives aux DOD.

L'estimateur de la variance des estimations du trafic international par paires d'origines et de destinations est une modification simple de l'estimateur de la variance utilisé dans l'enquête sur l'origine et la destination des passagers payants.

### 7.1 Variance de l'estimation du trafic avec correspondance

Le calcul de l'estimateur de la variance des estimations de l'origine et de la destination des passagers payants dépend du nombre de passagers que chaque billet attribue aux paires d'origines et de destinations (c'est-à-dire au domaine d'intérêt). Rappelons que chaque billet est choisi avec une probabilité de 0.1 et que l'itinéraire peut être divisé en plusieurs segments ou voyages DOD. Un billet peut faire inscrire 0, 1, 2, etc. passagers dans un domaine d'intérêt donné. Par exemple, l'itinéraire

YWG - AC - LON - BA - YYZ

serait coupé, selon le programme qui calcule les points de division, en deux voyages simples :

YWG - AC - LON  
et  
LON - BA - YYZ

croient que de telles erreurs dans cette enquête sont petites. Toutefois, des consultations sont actuellement en cours avec les grands transporteurs aériens canadiens pour trouver des façons de réduire les exigences des organismes gouvernementaux en matière de déclarations. Au terme de ces discussions, l'activité aéroportuaire pourrait devenir l'objet d'une enquête-échantillon. Lors-que viendra le moment d'établir le plan de sondage, il faudra tenir compte de l'importance d'obtenir des données précises sur le volume des passagers aux points d'entrée ou de sortie.

La technique d'attribution suppose que tout le trafic des passagers non-dénombrés est pris en compte dans le trafic sans correspondance. Une façon intéressante de vérifier cette hypothèse serait de comparer, pour une même période de référence, le nombre de passagers d'un transporteur canadien déclaré dans l'enquête sur l'activité aéroportuaire et une estimation faite à partir de l'échantillon de l'enquête sur l'origine et la destination des voyageurs DOD comportent le même transporteur pour le trajet international et le même aéroport canadien comme point d'entrée ou de sortie. Il faudrait pouvoir dire si les écarts entre le chiffre déclaré et les estimations proviennent de différences entre les notions appliquées par ces enquêtes et, dans l'affirmative, vérifier si ces différences sont prises en considération dans le système d'estimation de l'ASIPOD. Si le système n'en tient pas compte, cela pourrait confirmer que l'hypothèse selon laquelle tout le trafic des passagers non-dénombrés est compris dans le trafic sans correspondance soulève des problèmes.

L'étude d'un bon nombre de ces lacunes devrait donc être poussée. Le système actuel comporte de sérieuses lacunes, mais on n'a pas encore réussi à trouver une solution de rechange à la technique d'attribution, étant donné les contraintes de temps et de coût. Néanmoins, les améliorations apportées aux estimations du trafic pour chaque marché par le système d'estimation de l'ASIPOD seront appréciables.

La technique d'attribution est fondée sur le postulat que les voyages DOD tronqués représentent bien le trafic sans correspondance. Examinons l'exemple hypothétique que nous venons d'illustrer afin de voir si cette hypothèse est, par intuition, raisonnable. La technique d'attribution suppose que les passagers qui partent de Montréal à bord d'un vol d'un transporteur étranger partiront auront la même destination finale que les passagers dont l'origine est Winnipeg ou Toronto et qui passent par Montréal. Cela revient à utiliser le comportement des voyageurs des communautés ethniques, comme les groupes polonais de Toronto et allemands de Winnipeg, pour tracer le comportement des communautés à prédominance française de Montréal. En outre, le caractère douteux du postulat de similitude entre les voyages avec et sans correspondance a été démontré empiriquement dans une étude-pilote (Rosen et Conroy (1977)) effectuée par la Commission canadienne des transports, en 1977. Bref, il existe des arguments non seulement intuitifs, mais aussi empiriques qui réfutent le postulat de base de cette technique.

La précision des estimations du nombre de passagers par paire d'origines et de destinations est mise en cause par toute exception à l'hypothèse que les voyages DOD tronqués représentent bien le trafic sans correspondance. Or, comme dans le cas du tableau 1, de grands volumes de passagers sont attribués à des paires d'origines et de destinations à partir d'un échantillon "réel" de petite taille. Par exemple, la fraction d'échantillonnage "réelle" du marché formé entre le Canada et la région no 7 n'est pas 10%, comme dans l'enquête sur le trafic intérieur, mais 5.7% (c'est-à-dire  $(100\% - 42.7\%) \times 10\%$ ), à cause de l'exclusion du trafic sans correspondance de l'univers. Par conséquent, un échantillon de moins de 10% sert à répartir un grand volume de passagers. L'existence d'exceptions à l'hypothèse en question pourrait donc causer de grands biais dans les estimations.

Comme les chiffres tirés de l'enquête sur l'activité aéroportuaire constituent des repères sur le volume du trafic, leur précision est très importante. Bien qu'aucune évaluation n'ait été entreprise concernant l'ampleur du biais attribuable à des erreurs d'observation dans les chiffres sur l'activité aéroportuaire, les économistes qui analysent les statistiques sur l'aviation



<u>Nombre de passagers</u>	<u>dans l'échantillon</u>	<u>Estimation du nombre</u>	<u>de passagers</u>	<u>DOD</u>
----------------------------	---------------------------	-----------------------------	---------------------	------------

1	10	YWG-AC-YMX-BA-LON-LO-WAW
2	20	YYZ-CP-YMX-BA-LON-LH-HAM

où les codes doivent être interprétés comme suit:

<u>Code</u>	<u>Désignation</u>
YWG	Winnipeg
YMX	Montréal (Mirabel)
LON	Londres
WAW	Varsovie
YYZ	Toronto
HAM	Hambourg
AC	Air Canada
BA	British Airways
LO	LOT
CP	CP Air
LH	Lufthansa

Ainsi,  $D_i = 3$  et  $A_i = 120 - (1/.1) \times 3 = 90$ .

Les voyages DOD tronqués, la proportion de ces voyages inclus dans  $D_i$ , le nombre de passagers à réparer et les estimations totales pour le marché sont donc:

<u>Voyages DOD tronqués</u>	<u>Proportions</u>	<u>Passagers à</u>	<u>réparer (<math>a_j</math>)</u>
-----------------------------	--------------------	--------------------	-----------------------------------

YMX - BA - LON - LO - WAW	1/3	30
YMX - BA - LON - LH - HAM	2/3	60

#### 6.4 Lacunes de la technique d'attribution

Cette technique comporte toutefois certaines lacunes.



$$f_j = \frac{d_j}{(d_j/f) + a_j} \quad (1)$$

et où  $f_j$  est la fraction d'échantillonnage corrigée pour la  $j^e$  paire d'origines et de destinations dans l'enquête sur l'origine et la destination des passagers payants;

$d_j$  est le nombre de passagers des vols internationaux pour la  $j^e$  paire d'origines et de destinations dans l'échantillon de l'enquête sur l'origine et la destination des passagers payants, et

$a_j$  est le nombre de passagers attribués à la  $j^e$  paire d'origines et de destinations.

Il convient de souligner que, conformément au point (iii) ci-dessus,  $a_j$  peut s'exprimer:

$$\begin{aligned} d_j &= \sum_{i \in I} a_{ji} \\ a_j &= \frac{D_j}{D_i} \times A_i, \quad j \in I \end{aligned}$$

$$\text{où } D_j = \sum_{i \in I} d_{ji}$$

$$\text{et } A_i = \sum_{j \in I} a_{ji}$$

### 6.3 Un exemple d'application de la technique d'attribution

Prenons un exemple simple pour illustrer comment fonctionne la technique d'attribution. Les données sur l'activité aéroportuaire fournies par British Airways révèlent que 120 passagers sont embarqués à Montréal à destination de Londres. Dans ce cas,  $C_i = 120$ . Supposons que l'enquête sur l'origine et la destination des voyageurs payants n'indique que deux suites de voyages DOD:

dénombrée de la population cible, sont appelées des voyages DOD sans correspondance, puisqu'il s'agit de paires d'origines et de destinations de voyages simples assurés par des transporteurs étrangers qui n'ont pas de correspondance avec un transporteur participant canadien. Selon la technique d'attribution, on impute les voyages DOD sans correspondance au 1<sup>er</sup> groupe de la façon suivante :

(i) Une liste de tous les voyages DOD inclus dans  $D_i$  est dressée (c'est-à-dire au même aéroport, par le même transporteur étranger et dans la même direction).

(ii) Les trajets intérieurs sont éliminés (c'est-à-dire les trajets entre un point canadien et un point d'entrée ou de sortie canadien). (Ces trajets étant faits par un transporteur canadien, ils sont pris en compte dans l'enquête sur l'origine et la destination des passagers payants. Les voyages ainsi "tronqués" sont alors sans correspondance.)

(iii) Le volume des passagers non dénombrés,  $A_i$ , est réparti parmi les voyages DOD ainsi "échantillonnés", selon la proportion de ces voyages dans  $D_i$ . De nouveaux taux de passagers classés par origine et destination de voyages simples sont produits, et il s'agit alors de "passagers répartis".

La technique d'attribution repose sur l'hypothèse que les voyages DOD tronqués représentent bien le trafic sans correspondance, ce qui donne plus de poids à certaines paires d'origines et de destinations que dans l'enquête qui porte seulement sur l'origine et la destination des passagers payants. On peut donc trouver l'estimateur  $\hat{d}_j^i$ , qui correspond au nombre total de passagers observés sur le marché pour la  $j^{\text{e}}$  paire d'origines et de destinations, en corrigeant la fraction d'échantillonnage de la manière suivante :

$$\hat{d}_j^i = (1/f_j) \times d_j$$

g est le nombre d'aéroports canadiens qui servent de point d'entrée ou de sortie, et

$n_i$  est le nombre de transporteurs étrangers (non américains) détenant un permis d'atterrissage, qui desservent le  $i^e$  aéroport canadien de point d'entrée ou de sortie.

L'enquête sur l'origine et la destination des passagers payants permet, par recoupement des transporteurs et des aéroports canadiens, d'obtenir un classement semblable des passagers entrants et sortants par origine et destination du voyage simple. Ainsi, l'échantillon de cette enquête fournit également b comptages.

A la première étape de la technique d'attribution, le volume de passagers non dénombrés,  $A_i$ , inclus dans le groupe d'attribution  $i$ , peut être estimé au moyen de l'équation suivante:

$$A_i = C_i - (1/f) \times D_i \quad (i = 1, \dots, b)$$

où  $C_i$  est le volume total des passagers dans le groupe  $i$ , selon l'enquête sur l'activité aéroportuaire;

$f$  est la fraction d'échantillonnage de l'enquête sur l'origine et la destination des passagers payants (c'est-à-dire  $1/10$ );

$D_i$  est le nombre échantillonné de passagers des vols internationaux dans le groupe  $i$ ;

$A_i$  correspond ainsi à l'estimation du nombre de passagers déplacés par un transporteur étranger et qui tombent dans le  $i^e$  des b groupes d'attribution, mais pour lesquels il manque des renseignements sur l'origine et la destination.

A l'étape suivante, on attribue le volume des passagers non dénombrés à des paires d'origines et de destinations. Ces paires reliées à la partie non

où le "2" s'explique par le fait qu'il y a deux comptages, un pour les passagers qui entrent au Canada et un pour ceux qui sortent du pays,

$$b = \sum_{i=1}^g n_i$$

A partir de l'enquête sur l'activité aéroportuaire, on effectue b comptages repères des passagers entrants et sortants aux aéroports canadiens, par transporteur. La valeur de b, le nombre de catégories auxquelles on peut attribuer un volume de voyageurs, se définit de la façon suivante:

La technique d'attribution consiste d'abord à estimer le volume de passagers non dénombrés, pour ensuite répartir ce volume dans des paires d'origines et de destinations.

## 6.2 La technique d'attribution

monde, on obtiendrait de plus hauts pourcentages de non-dénombrement pour chaque région du monde. Donc, il semble que, de façon générale, plus les régions géographiques sont désagrégées, plus le pourcentage de non-dénombrement est élevé. Pour citer un exemple, le pourcentage de non-dénombrement des passagers qui ont voyagé entre l'Est du Canada et la région no. 7 est de 55% contre 42.7% pour l'ensemble du Canada (tableau 1). Afin de clarifier l'affirmation qu'un haut degré d'aggrégation géographique dans le choix des paires d'origines et de destinations produit généralement un faible pourcentage de non-dénombrement, il faut souligner que le non-dénombrement existe seulement dans le cas des vols des transporteurs étrangers qui prennent fin à un point d'entrée au Canada. Le dénombrement du trafic de passagers est exhaustif lorsque la partie canadienne des paires d'origines et de destinations n'est ni un point d'entrée, ni un point de sortie. Ainsi, lorsque le degré d'aggrégation géographique augmente, on inclut plus de vols avec correspondance, ce qui fait diminuer le pourcentage de non-dénombrement.

problème pour le système d'estimation de l'ASIPOD à cause de l'absence de données sur l'origine et la destination de la partie non dénombrée de la population cible.

Le tableau suivant donne un indice du volume de passagers qui se déplacent entre le Canada et neuf régions du monde. (Ces chiffres sont des estimations annuelles pour 1979. Les neuf régions du monde ne sont pas nommées parce que ces données sont confidentielles.)

Tableau 1 - Passagers des vols internationaux - estimations, 1979

Entre le Canada et...	Origine et destination des passagers payants	Non-dénombrement (le pourcentage du total est entre parenthèses)	Total
-----------------------	--	--	-------

La région no 1	116,050	495	(0.4)	116,545
La région no 2	325,200	2,020	(0.6)	327,220
La région no 3	99,010	12,628	(11.3)	111,638
La région no 4	67,200	14,558	(17.8)	81,758
La région no 5	575,010	126,076	(18.2)	703,086
La région no 6	205,180	101,464	(33.1)	306,644
La région no 7	1,221,040	908,876	(42.7)	2,129,916
La région no 8	54,410	56,807	(51.1)	111,217
La région no 9	45,330	57,267	(55.8)	102,597
Total mondial	2,708,430	1,282,191	(32.1)	3,990,621

D'après les pourcentages de non-dénombrement, il est évident que ce problème est important.

Si ce tableau était fait pour l'Est du Canada et les neuf mêmes régions du



contrôle de ces erreurs d'observation, erreurs peu faciles à corriger, n'est pas un objectif de ce programme de révision.

(ii) Mettre au point un système de calcul simple, facile à utiliser et qui fournit des messages indicateurs sommaires.

Quelque 625,000 enregistrements d'origine et de destination, et quelque 820,000 enregistrements concernant l'activité aéroportuaire doivent être traités annuellement avec un minimum d'intervention manuelle. Etant donné qu'une multitude de passagers est répartie parmi un nombre considérable de paires d'origines et de destinations, les messages imprimés à chaque étape du système doivent résumer le traitement jusqu'à ce point, tout en indiquant les problèmes éventuels.

(iii) Fournir des tableaux des estimations du nombre de passagers sur les vols internationaux, soit régulièrement ou au besoin, de manière à simplifier les analyses que les utilisateurs doivent effectuer.

(iv) Produire des estimations, dont la fiabilité est mesurable, sur l'origine et la destination des passagers à bord des vols internationaux à horaire fixe.

Bien qu'on ait cru dans le passé que la fiabilité de ces statistiques était variable, elle était de fait inconnue jusqu'à présent. L'estimation de la fiabilité de ces données permettra de mieux connaître la force des conclusions qu'on peut tirer de leur analyse. Les deductions faites sans une mesure de la fiabilité des données peuvent être en fait fort trompeuses.

## 6. RÉSOLUTION DU PROBLÈME DE NON-EXHAUSTIVITÉ

### 6.1 Importance du problème

Comme dans d'autres enquêtes, la non-exhaustivité du dénombrement pose un

Supposons également que deux suites de voyages simples sont possibles:

Winnipeg - Air Canada - Toronto - Air France - Montréal - Paris

Toronto - Air France - Montréal - Paris - British Airways -

Londres

Le point de sortie canadien dans cet exemple serait Toronto, puisque c'est là que les passagers empruntent le transporteur étranger.

## 5. OBJECTIFS DE LA RÉVISION

Les quatre principaux objectifs visés par le projet de révision, dont le quatrième sera analysé en détail dans ce document, sont les suivants:

(i) Éliminer les problèmes signalés dans le système actuel.

Comme le système actuel ne prévoit pas d'imputation dans le cas des codes de transporteur inscrits de façon illisible sur le volet des billets d'avion, la catégorie "transporteurs inconnus" devient la troisième en importance dans les totalisations. En outre, il n'y a aucune vérification des codes qui permette de déterminer si un transporteur vole en direction ou à partir d'aéroports où il est effectivement autorisé à atterrir. Compte tenu des besoins des totalisations actuelles, il faut donc prévoir la vérification et l'imputation des données illisibles ou fausses inscrites sur les volets internationaux, en plus des corrections nécessaires aux coupons de vol exclusivement intérieurs.

D'autres erreurs d'observation ont été relevées, mais il est difficile de les corriger à l'aide d'un système d'estimation. Ce type d'erreurs inclut les cas où les transporteurs interprètent mal les instructions relatives aux coupons de vol; des erreurs fréquentes dans les numéros d'ordre des billets, qui sont utilisés pour le choix de l'échantillon; des erreurs dans les systèmes de traitement des transporteurs, etc. Le

américains) qui exploitent des services internationaux à horaire fixe à destination ou en provenance du Canada. Depuis le 1er janvier 1982, 10 transporteurs américains et 21 transporteurs étrangers remplissent une déclaration pour chaque aéroport canadien qu'ils desservent. Tout nouveau transporteur étranger qui obtient un permis d'exploitation de services à horaire fixe au Canada est automatiquement inclus dans le système de collecte des données sur l'activité aéroportuaire.

Les données sur le flux de passagers dans l'ensemble du trafic aérien sont extraites des renseignements sur l'activité aéroportuaire. Par "flux de passagers", on entend le nombre de personnes qui, pendant une période déterminée, voyagent à bord d'un transporteur donné entre un aéroport canadien déclarant et un point adjacent, c'est-à-dire le dernier aéroport visité ou le prochain point prévu à l'itinéraire. La technique d'attribution ne tient compte que des déclarations fournies par les transporteurs étrangers (non américains). Les éléments d'information extraits de cette enquête, qui sont utilisés pour déterminer les marchés formés par des paires d'origines et de destinations, sont le nombre de passagers payants embarqués ou débarqués au Canada, le transporteur au point d'entrée ou de sortie canadien et le nom du point d'entrée ou de sortie canadien. Par "point d'entrée ou de sortie canadien", nous entendons un aéroport qui participe à l'enquête et où un transporteur étranger (non américain) arrive au Canada ou en sort.

Cependant, d'après l'enquête sur l'origine et la destination des passagers payants, le point d'entrée/de sortie canadien pour les transporteurs canadiens et américains est le premier/dernier point canadien dans un itinéraire pour les vols à destination/en provenance du Canada. Quant aux transporteurs étrangers, le point d'entrée ou de sortie correspond au point à l'intérieur du Canada où un passager emprunte les services d'un transporteur étranger ou s'en sépare. Prenons l'exemple suivant:

Supposons qu'Air France assure un service Toronto - Montréal - Paris, et qu'il y a embarquement de certains passagers à Toronto et d'autres à Montréal.

Les données sur l'activité aéroportuaire sont envoyées tous les mois par les transporteurs transcontinentaux (Air Canada et CP Air) et régionaux (Eastern Provincial, Québécois, Nordair et Pacific Western Airlines) du Canada, par Norcanair et par tous les transporteurs étrangers (y compris les transporteurs

- Le nombre de débarquements et d'embarquements de passagers payants.

- Le nom du dernier aéroport visité, dans le cas des arrivées, ou le prochain aéroport prévu à l'itinéraire, dans le cas des départs;
- Le point d'origine et la destination finale prévue pour le vol;

- L'aéroport déclarant;

- Le transporteur participant;

Les données sur l'activité aéroportuaire proviennent de l'état 32. Les caractéristiques relevées pour chaque vol incluent:

Le programme jusqu'à ce qu'un nouveau point de division ne puisse être trouvé. Lieu d'origine. Chaque voyage DOD ainsi formé est présenté de nouveau au règle générale, on divise les itinéraires à partir du point le plus éloigné du de comparaisons entre ces distances et la distance totale de l'itinéraire. En comporte le calcul de la distance entre divers points d'un itinéraire, suivi les statistiques sur l'origine et la destination des passagers aériens, et il des voyages complets. Ce procédé est automatisé dans le système qui regroupe points de division est appliqué aux données sur l'origine et la destination sens unique. Pour obtenir les voyages DOD, un programme qui calcule des retour comme les itinéraires symétriques et circulaires, en voyages simples à décompose les itinéraires semi-circulaires ("open-jaw"), et les itinéraires de voyage simple correspond à une seule direction. Cette méthode exige qu'on (DOD) sont les parties d'un itinéraire, lequel est divisé de façon que chaque ordre qui indique la direction suivie dans l'itinéraire". Les voyages simples défini comme "une suite de points de départ et de destination énumérés dans un



transports aériens de la Commission canadienne des transports, en collaboration avec le ministère des Transports. Les données sont recueillies auprès des transporteurs aériens, pour le compte du Comité des transports aériens, par le Centre des statistiques de l'aviation (CSA) de Statistique Canada. En vertu des Règlements sur les transports aériens de la Loi sur l'aéronautique, les transporteurs sont obligés de remplir les états (questionnaires) du CSA.

Les statistiques sur l'origine et la destination (O & D) des passagers payants parviennent au CSA grâce à l'état 35. Parmi les renseignements recueillis, on trouve entre autres:

- l'origine et la destination inscrites sur le billet;
- Les points où un passager s'arrête pour prendre une correspondance sur la même ligne ou sur une ligne différente;
- le transporteur inscrit sur chaque volet.

Les renseignements sur l'origine et la destination des passagers payants sont fournis chaque mois par les principaux transporteurs canadiens qui émettent des billets à tarif unitaire et exploitent des services à horaire fixe. Depuis le 1er janvier 1982, les sept transporteurs canadiens qui participent à cette enquête sont Air Canada, CP Air, Eastern Provincial Airways, Nordair, Pacific Western Airlines, Air Ontario et Québecair. Les données américaines sont recueillies par le Civil Aeronautics Board auprès de tous les transporteurs autorisés, à l'exception de ceux qui assurent les transports par hélicoptère et les transports intérieurs en Alaska. Les chiffres pour chaque trimestre sont regroupés et les renseignements en double supprimés de façon à constituer un fichier des itinéraires complets, d'après l'origine et la destination du voyage complet ("TOD", Ticket Origin and Destination).

Par contre, les statistiques sur l'origine et la destination des passagers aériens sont traitées selon le concept de l'origine et de la destination du voyage simple ("DOD", Directional Origin and Destination), qui peut être



sans correspondance qui le conduit de Paris à Montréal par Air France et de nouveau à Paris avec ce même transporteur, ce voyage ne figurera pas dans l'enquête sur l'origine et la destination des passagers payants. Toutefois, ce genre d'itinéraire fait effectivement partie de la population cible de l'étude des voyages internationaux.

Ce dénombrement incomplet de la population cible constitue en fait un problème de non-exhaustivité pour le système d'estimation de l'ASIPOD. Il s'agit d'un problème de "non-dénombrement" plutôt que de "sous-dénombrement" parce qu'il est impossible d'inclure une grande partie de l'univers dans le cadre de l'enquête sur le trafic international.

Dans le système actuel, on utilise les données des enquêtes sur l'origine et la destination des passagers payants et sur l'activité aéroportuaire, et on applique une méthode appelée la technique d'attribution, afin de produire des estimations globales pour chaque marché.

L'enquête sur l'activité aéroportuaire procède à un dénombrement complet des passagers à bord des vols internationaux à destination ou en provenance de chaque aéroport canadien. Tous les transporteurs canadiens, américains et étrangers qui effectuent des vols à horaire fixe sont inclus, mais sans considération du premier point de départ ou de la destination finale des passagers. Ainsi, l'enquête sur l'activité aéroportuaire donne un dénombrement du volume total des passagers de tous les transporteurs qui desservent la population cible. La technique d'attribution consiste à estimer le volume de passagers non dénombrés par l'enquête sur l'origine et la destination des passagers payants et à les classer dans des paires d'origines et de destinations. Avant d'expliquer plus en détail cette technique, il convient d'abord de décrire ces deux enquêtes sur le trafic aérien de façon un peu plus complète.

#### 4. PRINCIPAUX ASPECTS DES DEUX ENQUÊTES

L'organisation de ces deux programmes d'enquête est confiée au Comité des

destination des passagers entre le Canada et les États-Unis, dans lequel les États-Unis fournissent au Canada des renseignements sur les itinéraires complets où:

- (i) un point situé au Canada et un point situé aux États-Unis figurent sur le billet, ou
- (ii) un transporteur des États-Unis s'est rendu en un point situé au Canada ou a quitté un point canadien, ou
- (iii) un transporteur canadien s'est rendu en un point situé aux États-Unis ou a quitté un point américain.

A cause de cet accord, nous employons les expressions "étranger" ou "américain" pour désigner ce qui est "ni canadien, ni américain".

Ce type de billets est également l'univers de l'enquête sur l'origine et la destination des passagers payants, mais celle-ci vise uniquement les principaux transporteurs canadiens. Chaque transporteur participant prend un coupon des billets dont le numéro d'ordre se termine par "0", tenant compte seulement des volets dont il est le premier à prélever. Ainsi, l'enquête porte sur un échantillon de 10% des coupons de vol pour des itinéraires qui incluent les services d'au moins un des transporteurs canadiens ou américains participants.

Quelques renseignements sur les marchés des transporteurs étrangers (non américains) peuvent être tirés de l'enquête sur l'origine et la destination des passagers payants. Par exemple, si un passager voyage par Air Canada d'Ot-tawa à Montréal, et de là à Paris par Air France, l'enquête sur l'origine et la destination des passagers payants comptera ce trajet international parce qu'un transporteur canadien intervient quelque part dans l'itinéraire. Le transporteur canadien doit déclarer tous les détails sur les transporteurs et les trajets, y compris la partie concernant Air France.

L'enquête sur l'origine et la destination des passagers payants ne tient pas compte cependant des volets où des transporteurs étrangers sont inscrits du début à la fin d'un itinéraire. Par exemple, si un passager suit un itinéraire

3. Obtenues à partir de totalisations internes de la Commission canadienne des transports.

La population cible de ce système est l'ensemble des billets émis pour un voyage international (c'est-à-dire entre un pays étranger et le Canada ou les Etats-Unis). Il existe un programme d'échange de données sur l'origine et la non celui des estimations de la fiabilité de ces chiffres.

Le calcul des estimations du nombre de passagers des vols internationaux, mais le maintien de la méthodologie fondamentale a fourni les grandes lignes du miner lesquelles pouvaient être adoptées compte tenu des délais et du budget. L'origine et de la destination souhaitables, à les classer selon leur degré d'utilité et à déterminer. Par conséquent, la plupart de ces travaux ont visé à chercher des améliorations. Il faut conserver la méthodologie fondamentale appliquée dans le système Le remaniement effectué dans le cadre de ce projet repose sur le principe

### 3. NATURE DU PROBLÈME DE NON-EXHAUSTIVITÉ

Liards de dollars canadiens en 1980. naux à destination ou en provenance du Canada ont atteint près de 2,3 milliards de dollars canadiens en 1980. mentionnons que les recettes<sup>3</sup> de l'ensemble des services aériens internationaux. A titre indicatif de l'importance de ces données pour l'économie canadienne, l'origine et de la destination des passagers à bord des vols internationaux. une juste part du marché, il est essentiel de posséder des estimations de transporteurs canadiens et étrangers. Afin de permettre au Canada de négocier pays, qui découlent de l'attribution de diverses routes internationales à des d'analyses de coûts et de bénéfices, tant pour le Canada que pour les autres s'ils sont en expansion. En conséquence, ils doivent disposer des résultats ciateurs doivent savoir où se trouvent les marchés du transport aérien et Pour discuter de ces sujets et surtout pour échanger des routes, les négoc-

- l'exemption d'impôts sur le revenu.

- les transferts de fonds;

genre de service s'appelle un service à tarif unitaire.

Avant qu'un service commercial international à horaire fixe à destination ou en provenance du Canada puisse être assuré, il doit exister un accord officiel entre le Gouvernement du Canada et le gouvernement du deuxième pays. Cet accord peut prendre la forme d'un échange provisoire de notes diplomatiques ou d'une entente globale sur le transport aérien conclue par voie de négociations.

Des négociations bilatérales sur les transports aériens nécessitent la participation de fonctionnaires des ministères canadiens des Affaires extérieures, des Transports et de l'Industrie et du Commerce, ainsi que de la Commission canadienne des transports. Les pourparlers peuvent durer quelques mois ou quelques années.

Les routes aériennes pour les services à horaire fixe sont normalement l'objet principal des négociations, mais non le seul. Les articles contenus dans une entente sur le transport aérien peuvent concerner :

- Le droit de survoler un certain territoire, ou d'y atterrir pour des raisons non liées au service normal;

- La désignation du transporteur qui peut desservir chaque route;

- Le respect des lois et des règlements de chaque pays en matière d'entrée, d'autorisation, d'immigration, de passeports, de douanes et de quarantaine;

- La navigabilité, les certificats de compétence et les permis;

- L'échange de statistiques;

- Les tarifs;



estimations du nombre de passagers sur les vols internationaux à horaire fixe, entre le Canada et les marchés étrangers, en fonction de diverses paires d'origines et de destinations. La première source, l'enquête sur l'origine et la destination des passagers payants, offre un échantillon de données sur les voyages internationaux où au moins un parcours d'un itinéraire est assuré par un transporteur canadien. Cette enquête, toutefois, présente des lacunes au niveau du champ d'observation, car elle ne fournit pas de données sur les voyages internationaux qui comportent les services d'un transporteur étranger pour l'ensemble de l'itinéraire. La deuxième source, l'enquête sur l'activité aéroportuaire, dénombre tous les passagers qui sortent du Canada ou qui y entrent d'un autre pays à bord d'un transporteur canadien ou étranger qui offre des services à horaire fixe, sans tenir compte de l'origine ou de la destination des passagers.

Ce document présente d'abord un aperçu des besoins des utilisateurs en matière de données sur l'origine et la destination des passagers sur les vols internationaux. On examine ensuite les principaux aspects des deux sources de données sur le trafic aérien, ainsi que le problème de la non-exhaustivité du dénombrement. Enfin, ce texte explique de quelle façon le système d'estimation de l'ASIPOD permet de calculer des estimations du nombre de passagers, et les coefficients de variation correspondants, pour diverses paires d'origines et de destinations, que ces villes soient situées ou non dans les parties du marché international visées par l'enquête sur les passagers payants.

## 2. BESOIN DES UTILISATEURS

Des estimations par origine et destination des passagers des vols commerciaux internationaux à horaire fixe sont nécessaires aux utilisateurs pour les négociations bilatérales.

Un service aérien commercial international à horaire fixe est défini comme l'exploitation de vols entre des points situés au Canada et des points à l'extérieur pour le transport de passagers, de biens ou de courrier par aéronef selon un horaire fixe et moyennant un taux de transport unitaire. Ce



# MÉTHODOLOGIE DU SYSTÈME CANADIEN D'ESTIMATION DE L'ORIGINE ET DE LA DESTINATION DES PASSAGERS DES VOLS INTERNATIONAUX À HORAIRE FIXE<sup>1</sup>

Greg Hunter et Lisa DiPietro<sup>2</sup>

Le système d'estimation de l'origine et de la destination des passagers des vols internationaux à horaire fixe, communément appelé l'ASIPOD (Air Scheduled International Passenger Origin and Destination) est fondé sur les données de deux enquêtes sur le trafic aérien et permet de produire des estimations du nombre de passagers à bord des vols internationaux en fonction de l'origine et de la destination. La technique d'attribution offre une solution au problème du non-dénombrément du trafic sans correspondance. Les hypothèses sur lesquelles reposent cette technique soulèvent assez de doutes pour qu'une évaluation du biais des estimations mérite d'être entreprise. Cependant, le nouveau système comportera des améliorations importantes qui diminuent le biais des estimations, tout en calculant un estimateur de la fiabilité des résultats. Par conséquent, les négociations bilatérales sur le transport aérien international nous permettront de mieux déterminer la valeur des conclusions tirées au sujet des marchés des transports aériens à partir de ces estimations.

## 1. INTRODUCTION

En 1979, Statistique Canada a entamé une révision du programme fédéral de la statistique de l'aviation en invitant le ministère des Transports et la Commission canadienne des transports (les deux "organismes utilisateurs") à former une équipe interministérielle. Le remaniement du système d'estimation de l'ASIPOD est un des divers projets de révision du programme.

Selon le système d'estimation de l'origine et de la destination des passagers des vols internationaux à horaire fixe, communément appelé l'ASIPOD, on puise les données de deux enquêtes sur le trafic aérien pour produire des

<sup>1</sup> Présenté aux réunions conjointes de l'American Statistical Association à Cincinnati, août 1982.

<sup>2</sup> Greg Hunter, Division des méthodes d'enquêtes-entreprises, Statistique Canada. Lisa DiPietro, Division des transports et des communications, Statistique Canada.

- [3] "Report of the Interdepartmental Working Group on For-hire Trucking - Phase II Review"; rapport rédigé par la Division des transports et des communications de Statistique Canada, avril 1979.

- [4] Statistique Canada, Entreprises de camionnage et de déménagement, no 53-222 au catalogue, annuel.

- [5] Lussier, R. (1981), "For-hire Trucking Survey: Survey Design", Techniques d'enquête, Statistique Canada, vol. 7, no 1, pp. 74-92.

Faut procéder à des analyses pour évaluer les renseignements fournis. En ce qui a trait à l'année de référence 1981, seulement une bande sera utilisée; elle porte sur un transporteur qui a effectué environ cinq millions d'expéditions en 1981. Au cours des prochaines années, plus d'entreprises fourniront des données sur bande, et des accords sont sur le point d'être conclus avec cinq autres transporteurs pour l'année de référence 1982.

Toutefois, lorsqu'on obtient finalement une bande de données qui satisfait aux exigences de Statistique Canada, on doit tout de même faire beaucoup de traitement informatique pour exploiter la bande. En outre, des interventions manuelles sont nécessaires pour régler le problème des données qui ne figurent pas sur les diverses listes. Il faudra donc probablement échantillonner les dossiers contenus sur chaque bande selon le plan de sondage du deuxième degré appliqué aux documents d'expédition des entreprises des classes 1 et 2.

Une autre possibilité qui devrait être explorée serait de demander aux entreprises elles-mêmes d'échantillonner leurs documents. Par exemple, au moment où les bordereaux d'envoi sont établis, un transporteur pourrait photocopier ceux dont le numéro de série se termine par un certain chiffre et faire parvenir ces photocopies à Statistique Canada tous les mois.

Enfin, des travaux seront consacrés à une évaluation complète des diverses parties de l'enquête et à la formulation de propositions pour l'améliorer. On espère pouvoir appliquer ces propositions à l'enquête portant sur l'année de référence 1982.

## BIBLIOGRAPHIE

[1] Statistique Canada, L'enquête sur le transport routier de marchandises pour compte d'autrui, no 53-224 au catalogue, annuel.

[2] "Report on the Findings and Recommendations of the Working Group on the For-hire Trucking Survey Phase I Review"; rapport rédigé par la Division des transports et des communications de Statistique Canada, en date du 7 juillet 1978.

Comme il est mentionné dans une section antérieure, l'enquête exploite trois sources de données, dont les bandes magnétiques fournies par certains répondants. Ce genre de données s'est avéré difficile à traiter et, malgré les travaux déjà commencés sur ce problème, les résultats sont décevants jusqu'à présent. Il faut entreprendre des négociations complexes avec les entreprises afin d'obtenir les données nécessaires sur bande et ensuite il

## 15. PROJETS

Premièrement, la publication relative à l'enquête présentera les estimations produites par le système régulier, de même que des mesures d'erreur comme les coefficients de variation. Deuxièmement, on répondra aux demandes individuelles, mais sous réserve des contraintes de coût et de fiabilité des renseignements. Enfin, la base des données sur les expéditions qui est constituée pour cette enquête pourrait être mise à la disposition de certains utilisateurs sur bandes magnétiques, à condition que les normes de confidentialité soient respectées.

L'enquête remaniée diffusera des renseignements de trois façons, comme le faisait l'ancienne enquête.

### 14.2 Méthodes de diffusion

Dans le passé, les demandes spéciales visant des estimations provenaient de ministères intéressés par le commerce, d'agences fédérales et provinciales de réglementation des transports, de transporteurs, d'experts-conseils d'universités, d'associations industrielles et d'un bon nombre d'autres organismes et particuliers qui s'intéressent au secteur des transports.

Une utilisation particulière de l'enquête a été de définir les caractéristiques des marchés du camionnage à l'aide de variables telles que la marchandise transportée, la longueur moyenne des parcours et le poids des expéditions. Une autre étude a été réalisée sur divers aspects des activités des transporteurs dans des domaines réglementés et non réglementés. Elle contient une analyse du comportement des transporteurs face aux coûts à partir de caractéristiques du trafic, comme la taille des expéditions et la longueur moyenne des parcours.



L'ensemble des données sur les entreprises des classes 1 et 2 est épuré par l'élimination des expéditions hors du champ de l'enquête. Ces dernières incluent les expéditions à destination ou en provenance des Etats-Unis, les expéditions qui totalisent 15 milles ou moins de l'origine à la destination, les expéditions faites hors des routes publiques, les expéditions qui seraient comptées deux fois parce qu'elles sont des expéditions de transfert ou parce qu'elles ont été déclarées par des démagueurs qui sont des agents de transporteurs sur longue distance et aussi par les transporteurs sur longue distance eux-mêmes, les expéditions qui n'ont pas produit de recettes au titre du transport interurbain, et les enregistrements concernant les services autres que le transport, comme l'entreposage, l'emballage, la location de matériel, le chargement et le déchargement.

Les estimations des recettes et du nombre de tonnes et de tonnes-kilomètres, qui doivent être publiées, sont enfin produites par la sommation des données pondérées dans les domaines d'intérêt correspondants. Des mesures d'erreur, comme les coefficients de variation, sont fournies avec les estimations. Les coefficients de variation sont calculés à l'aide de la formule issue du plan de sondage, mais on fait aussi l'hypothèse que l'échantillon systématique d'expéditions est un échantillon aléatoire simple.

#### 14. UTILISATION DES DONNÉES ET MÉTHODES DE DIFFUSION

##### 14.1 Utilisation des données

Les demandes d'estimations faites à partir de l'ancienne enquête émanent d'un grand nombre de sources. De plus, la nature de ces demandes est très variée. Il est prévu que la demande d'estimations faites à partir de la nouvelle enquête sera semblable à ce qu'elle était dans le passé.

Ces estimations sont devenues des instruments très répandus pour satisfaire cinq besoins principaux, à savoir mesurer le volume des marchandises du commerce intérieur transportées dans les provinces et entre celles-ci par les transporteurs interurbains pour compte d'autrui, déterminer le taux de croissance du transport interurbain de marchandises, fournir des données sur l'expansion régionale, aider à la réalisation d'études sur le transport et étayer la présentation de mémoires, de communications et de demandes aux organes et commissions de réglementation.



transcrits et à comparer ces chiffres avec le total du nombre d'expéditions, des recettes et du nombre de tonnes que le représentant du LDD a déclaré au cours de l'interview. Des vérifications semblables sont faites pour les entreprises de la classe 3. Dans tous les cas où on trouve un écart, on effectue un suivi.

### 13. TECHNIQUES D'ESTIMATION

Pour l'estimation, on a décidé de considérer l'échantillonnage systématique au deuxième degré dans les entreprises des classes 1 et 2 comme un échantillonnage aléatoire simple sans remise (EASSR). On a pris cette décision parce que, en premier lieu, on a considéré les documents comme étant en ordre aléatoire et, en deuxième lieu, un EASSR permet de calculer une estimation de la variance échantillonnale.

La première étape de l'estimation consiste à calculer des coefficients de pondération. Il y a des coefficients de premier et deuxième degré pour les expéditions des entreprises des classes 1 et 2, mais seulement des coefficients de premier degré en ce qui concerne les entreprises de la classe 3. Généralement, un coefficient de premier degré correspond à l'inverse de la probabilité qu'un LDD soit sélectionné dans sa strate, alors qu'un coefficient de deuxième degré représente l'inverse de la probabilité qu'une expédition d'un LDD soit choisie supposant qu'on effectue un EASSR. L'ordinateur corrige les coefficients de premier degré en fonction du nombre d'entreprises pour laquelle la collecte des données a échoué, mais aucune correction n'est faite pour les fermatures de LDD ou d'entreprises, ni pour les sorties d'entreprises du champ de l'enquête, étant donné qu'on les considère comme n'ayant fait aucune expédition. Un coefficient de pondération final est inscrit sur chaque enregistré dans l'ensemble de données des entreprises des classes 1 et 2 et dans l'ensemble de données des entreprises de la classe 3.

Des rapports détaillés sont produits sous la forme de tables contenant diverses totalisations des données. Ces chiffres sont très utiles à l'analyse des données et aux dernières vérifications de la qualité des données.

Lorsqu'une paire ne figure pas sur la liste, on calcule une distance aérienne (X) à l'aide de la latitude et de la longitude des points d'origine et de destination. Ensuite, X est convertie en distance routière Y au moyen du modèle de régression linéaire simple

$$Y = aX + b$$

où a et b varient en fonction de 12 régions d'origine et de 12 régions de destination. La distance routière est ajoutée aux autres données.

Il y a aussi imputation des renseignements qui manquent concernant les exportations partiellement transcrites des entreprises des classes 1 et 2. La technique d'imputation qu'on applique varie d'après la variable ou la paire de variables manquantes. Les principales imputations se font à l'aide de relations fixes entre les chiffres transcrits, de facteurs de conversion des unités de poids et de tables de distribution au prorata. Voici un exemple d'une relation fixe entre les renseignements déclarés:

$$\text{poids} = \frac{\text{recettes} \times 100}{\text{tarif}}$$

Cette formule peut servir à imputer le poids lorsque les recettes et le tarif sont connus ou à imputer les recettes quand le poids et le tarif figurent parmi les données. Les facteurs de conversion des unités de poids sont des coefficients déterminés par type d'unité (par exemple, caisse, sac, litre, etc.) et par code de la CTP. Sachant l'unité et le code de la CTP, on peut appliquer le bon facteur de conversion au nombre d'unités pour calculer le poids. Enfin, les tables de distribution au prorata indiquent le tarif par catégorie de marchandise, par bloc numérique de distance parcourue et par groupe de recettes ou de poids. Ces tables sont fondées sur les données de l'année précédente, mais sont mises à jour à mesure qu'on traite les données valides recueillies pour la période courante. Au moyen de ces tables, on calcule le poids lorsque les recettes sont connues, ou les recettes quand le poids est connu.

Dans les cas où trop de caractéristiques doivent être imputées, l'expédition est classée comme inutilisable.

Par la suite, des vérifications d'extrapolation sont effectuées. Dans le cas des entreprises de classe 1 ou 2, ce type de vérification consiste à pondérer approximativement le nombre d'expéditions, les recettes et le nombre de tonnes

Les autres traitements incluent la conversion du poids en unités métriques et du tarif en \$/100 kilogrammes. On procède également à une comparaison des noms des origines et des destinations (c'est-à-dire les villages, les villes, etc.) avec une liste de municipalités dans le but d'obtenir un code de la Classification géographique type (CGT), ainsi que la latitude et la longitude. Lorsque le nom de l'origine ou de la destination ne figure pas sur cette liste, l'opérateur doit utiliser un synonyme. De même, le nom de chaque marchandise est assorti à une liste de marchandises pour trouver un code (3 chiffres) de la Classification type des produits (CTP). Si la marchandise ne figure pas sur la liste, l'opérateur se sert d'un synonyme ou inscrit le mot "unknown". Il y a donc toujours un code de la CTP inscrit pour chaque expédition. De plus, le mini-ordinateur produit le nombre requis d'enregistrements de transcription pour chaque type d'expédition à partir des données tirées des profils des LDD des classes 1 et 2.

Enfin, les données sont extraites du mini-ordinateur et deux ensembles de données sont créés: un ensemble pour les expéditions des entreprises des classes 1 et 2 et un autre pour les renseignements concernant les types d'expéditions des entreprises de la classe 3. La principale différence entre ces deux ensembles de données est que le premier porte sur des expéditions distinctes tandis que le deuxième contient des données globales. Il convient aussi de noter que le premier compte plus de variables (par exemple, le tarif, le lieu d'origine plutôt que la province d'origine, etc.) que le deuxième ensemble.

### 12.3 Contrôles du système principal et imputations

La distance routière entre l'origine et la destination de chaque expédition des entreprises des classes 1 et 2, qui est visée par l'enquête, doit être connue afin d'estimer le nombre de tonnes-kilomètres parcourus par les transporteurs des classes 1 et 2. On vérifie donc chaque paire d'origines et de destinations, codées selon la CGT, en fonction d'une liste de distances pour obtenir la distance routière en kilomètres entre les deux points.

référence. À titre d'exemples, une entreprise pour laquelle on constate que 100 % de ses recettes proviennent d'expéditions locales est considérée comme hors du champ de l'enquête, tandis qu'un transporteur qui possède un seul LDD et qui refuse de collaborer ou est touché par une grève est considéré comme un cas d'échec.

Deuxièmement, il y a une vérification des données recueillies sur les profils des LDD des classes 1 et 2 afin de déterminer le nombre d'expéditions qui auraient dû être transcrites dans chaque catégorie d'expédition déclarée, si tous les documents avaient été dans les dossiers. On obtient ce nombre en faisant des calculs avec le nombre total d'expéditions comprises dans un profil, et avec l'origine choisie au hasard et l'intervalle de sondage qui auraient été utilisés si les documents avaient été disponibles. Le résultat est ensuite codé pour que l'ordinateur puisse produire le nombre requis d'enregistrements de transcriptions dans chaque catégorie d'expédition, comme si on avait obtenu originalement toutes les transcriptions.

## 12.2 Saisie des données

Après le traitement manuel, les formules arrivent au stade de la saisie des données. Cette opération est faite au moyen d'un mini-ordinateur qui permet d'effectuer des vérifications et d'autres traitements en ligne.

Diverses vérifications sont faites sur le mini-ordinateur. Certaines produisent des messages d'erreur et requièrent des corrections, mais d'autres font apparaître des avertissements qui demandent de vérifier les données introduites en ordinateur et d'apporter des corrections seulement s'il y a lieu. Certaines vérifications concernent la validité de chaque réponse individuellement tandis que d'autres examinent les rapports entre les caractéristiques valides d'une même expédition. Les opérateurs du mini-ordinateur sont censés connaître l'enquête en détail de façon à pouvoir apporter des corrections en ligne. Des imputations manuelles sont effectuées au besoin, étant donné qu'il n'y a pas d'imputation automatique pour les LDD de la classe 3.



#### 11.5 "Profils" des LDD des classes 1 et 2

Il arrive parfois qu'un LDD de la classe 1 ou 2 ne peut pas fournir de documents, ne conserve pas de documents utilisables pour l'échantillonnage ou ne peut pas fournir une partie de ses documents d'expédition et qu'il est impossible de représenter cette partie au moyen des documents disponibles. Ce dernier cas peut se produire, par exemple, lorsque les documents qui manquent correspondent à des contrats particuliers qui ont été enlevés des dossiers pour des besoins de vérification comptable. Dans ces cas, l'interviewer doit faire un "profil" des documents qui manquent, c'est-à-dire qu'il doit demander à un représentant d'un LDD de décrire les types d'expédition consignés sur les documents qui ne peuvent être consultés. Les profils ressemblent aux descriptions des types d'expédition établies pour les entreprises de la classe 3, sauf qu'il faut indiquer l'origine et la destination précises de chaque expédition (c'est-à-dire le village, la municipalité, la ville, etc.).

Dresser un profil peut prendre beaucoup de temps dans certains LDD parce que leurs activités peuvent être assez complexes. Il faut une bonne collaboration de la part des représentants des LDD.

## 12. TRAITEMENT DES DONNÉES

### 12.1 Traitement manuel

Les formules remplies sont envoyées au bureau central de Statistique Canada, à Ottawa, où elles sont enregistrées et où on vérifie les numéros d'identification. Deux petites tâches sont également accomplies avant l'étape suivante.

Premièrement, un bref examen est effectué afin de repérer et de coder les fermetures de LDD ou d'entreprises, la sortie d'une entreprise du champ de l'enquête et l'échec de la collecte des données auprès d'un transporteur. Une entreprise qui tombe hors du champ de l'enquête est un transporteur actif dont les recettes comprises dans le champ de l'enquête sont nulles pour l'année de référence. Une entreprise pour laquelle on dit que la collecte des données a échoué est un transporteur actif pour lequel aucune information n'a été recueillie bien que les interviewers savaient que ce transporteur avait des recettes non nulles comprises dans le champ de l'enquête pour l'année de



Une fois qu'une expédition est choisie, l'interviewer transcrit les caractéristiques requises. Ce travail est souvent difficile parce qu'il n'est pas toujours facile de comprendre les divers documents et les codes qui figurent sur certains. Ce problème s'applique surtout aux noms des marchandises, et l'interviewer doit éviter d'inscrire des marques de commerce, des noms propres et des noms qui ont plus d'une signification. Il faut souvent interpréter les renseignements consignés dans un document et transcrire les données sur des feuilles de codage sous une forme exploitable par l'ordinateur.

#### 11.4 Description générale de la collecte de données dans les entreprises de classe 3

L'interviewer envoie une lettre de présentation à l'entreprise deux ou trois semaines avant de téléphoner. Par la suite, l'interviewer communique avec le représentant de l'entreprise qui est le mieux placé pour fournir les renseignements requis, ce qui peut nécessiter plusieurs appels téléphoniques. Cela fait, l'interviewer procède à une interview téléphonique.

Les questions posées pendant l'interview sont semblables à celles posées dans le cas des LDD des classes 1 et 2, mais il y a une différence importante: aucune question ne porte sur le genre de documents utilisés ou sur les méthodes de classement appliquées par l'entreprise. Lorsque cette première partie de l'interview est terminée, l'interviewer demande au répondant de décrire les types d'expédition dont s'occupe son entreprise. Pour chaque catégorie d'expédition, la description doit spécifier la province d'origine, la province de destination et le nom de la marchandise transportée. On demande ensuite au représentant de donner une estimation du nombre d'expéditions, du poids moyen et des recettes moyennes provenant de chaque catégorie d'expédition.

Les experts-économistes en statistiques du transport considèrent généralement que les activités de l'entreprise de la classe 3 sont assez homogènes. Chaque transporteur a donc seulement un petit nombre de types d'expédition à déclarer. On croit aussi que la couverture obtenue par cette méthode est acceptable du point de vue des utilisateurs. Toutefois, aucun test n'a été appliqué à cette hypothèse concernant les transporteurs de la classe 3.

Les types de système de classement comprennent par ordre numérique absolu, par ordre numérique avec interruption, par ordre chronologique, par ordre alphabétique (par exemple suivant le nom du client), par terminus, par genre de marchandise, ou sans aucun ordre. Les documents peuvent même être recoupsés, par exemple, suivant le numéro d'ordre et d'après le nom du client. Dans un même système de classement, les documents peuvent être conservés dans des tiroirs de classeurs, dans des reliures ou des fichiers "Shannon", sur des rayons, dans des tiroirs ou même dans des livres.

Les données globales sur les activités du LD couvrent plusieurs variables, comme le total des recettes provenant du transport, le nombre total de tonnes transportées, le nombre total d'expéditions, le pourcentage de ces trois derniers éléments qui correspond aux expéditions interurbaines et aux expéditions internationales, les types de marchandises transportées et le pourcentage que chacun de ces types représente dans le total des recettes au titre du transport.

Il arrive souvent que l'intervieur puisse choisir parmi quelques méthodes de classement pour trouver les documents qui fournissent les renseignements requis. L'intervieur évalue à quel point les divers systèmes de classement sont complets par rapport aux cinq principales caractéristiques et l'année de référence, et il choisit le système qui présente le moins de sous-dénombrement. Toutefois, si le sous-dénombrement (le cas échéant) a la même ampleur dans deux systèmes ou plus, l'intervieur prend celui qui comprend le plus petit nombre de documents hors du champ d'enquête ou celui qui permet d'enlever ces documents du fichier ou de ne pas les compter.

L'intervieur sélectionne ensuite l'échantillon d'expéditions de la façon suivante. A partir du nombre d'expéditions déclaré par le représentant du LD, l'intervieur obtient d'une table un intervalle de sondage et une origine choisie au hasard. Dans certains cas, l'intervaleur et l'origine peuvent avoir été déterminés préalablement par le bureau central de Statistique Canada. Puis, lorsque les documents sont classés en ordre numérique, on additionne l'origine choisie au hasard ou l'intervaleur de sondage aux numéros d'ordre des documents pour obtenir les expéditions faisant partie de l'échantillon. Autrement, il faut compter un nombre de documents égal à l'origine choisie au hasard ou à l'intervaleur pour sélectionner les expéditions.

d'exercices présentés par le chargé d'enquête et un ou des spécialistes en méthodes d'enquête. Pendant la formation sur le terrain, des groupes de trois ou quatre personnes se rendent à un LDD afin d'appliquer les connaissances acquises en classe et afin d'en discuter.

## 11.2 Planification des travaux de collecte

Après leur formation, les chargés de projet des opérations régionales recrutent les interviewers et leurs donnent à leur tour une formation complète. Par la suite, les interviewers, suivant les conseils de leur chargé de projet, se fixent un horaire de travail et planifient leurs itinéraires de visites aux LDD des entreprises des classes 1 et 2. Les itinéraires sont élaborés de façon à éviter les déplacements inutiles et à assurer le meilleur rendement possible. Les interviewers envoient aux représentants des LDD une lettre de présentation qui décrit brièvement la nature de l'enquête. Plus tard, les interviewers communiquent par téléphone avec eux pour fixer un rendez-vous. La collecte des données a lieu entre mai et septembre pour l'enquête qui porte sur l'année civile précédente.

## 11.3 Description générale de la collecte des données dans les LDD des entreprises des classes 1 et 2

Au moment du rendez-vous, l'interviewer mène une interview avec les représentants des LDD, leur expliquant l'enquête et à quoi servent les données. L'interviewer évalue le temps nécessaire pour effectuer le travail et demande des renseignements au sujet de l'entreprise. Ces questions concernent principalement les modifications à apporter aux noms et aux adresses, les changements de propriétaire, le(s) type(s) de documents et le système de classement utilisés, et des données globales sur les activités du LDD pendant l'année de référence.

Les types de documents d'expédition les plus communs sont les bordereaux d'envoi, les connaissements, les manifestes, les feuilles de route et les factures. Une entreprise peut se servir de n'importe quelle combinaison de ces documents.

dangereuses, ou d'autres types d'entreprises pour lesquelles le chargé d'enquête pourrait vouloir élargir la base des données. Au cours des prochaines années, il sera peut-être aussi nécessaire de faire un ajustement pour les entreprises qui exercent leurs activités dans un domaine où la fiabilité des estimations obtenues l'année précédente atteignait plus ou moins le niveau souhaité.

Pour les LDD des entreprises de la classe 3, il n'y a pas de deuxième degré d'échantillonnage. On n'échantillonne pas des expéditions distinctes dans les dossiers des LDD, mais on recueille des données globales au niveau des LDD.

## 11. OPÉRATIONS SUR LE TERRAIN

Les opérations sur le terrain sont différentes pour les entreprises des classes 1 et 2 et celles de la classe 3. Dans le cas des deux premières classes, le travail consiste à tirer des documents d'expédition des dossiers des LDD et à transcrire les caractéristiques de ces expéditions sur des feuilles de codage. Pour les entreprises de la classe 3, on recueille par téléphone des renseignements globaux sur les activités du transporteur.

La présente section porte sur les tâches du personnel des opérations régionales de Statistique Canada, notamment la formation des chargés de projet pour les opérations régionales, la planification des travaux de collecte, la collecte de données aux LDD des entreprises des classes 1 et 2, la collecte de données auprès des transporteurs de la classe 3 et, enfin, les profils des LDD des classes 1 et 2.

### 11.1 Formation des chargés de projet pour les opérations régionales

Toutes les ans, les chargés de projet des opérations régionales de Statistique Canada reçoivent une formation concernant tous les aspects de l'enquête. Les séances de formation durent quatre jours et ont lieu au mois de mars. Ce programme est divisé en deux parties : une formation théorique et une formation sur le terrain. La formation théorique comprend une série d'exposés et



La sélection d'un échantillon systématique d'expéditions parmi les dossiers de chaque LDD choisi constitue le deuxième degré du plan d'échantillonnage des LDD des entreprises des classes 1 et 2. Le choix des documents est effectué dans les LDD par les interviewers de la Division des opérations régionales de Statistique Canada. L'intervalle de sondage varie en fonction du nombre d'expéditions faites par une entreprise et, généralement, les interviewers le trouvent dans une table de distribution qui leur est fournie. Cette table indique des intervalles de nombres possibles de documents d'expédition d'une compagnie et l'intervalle de sondage pour chaque cas. Cet intervalle peut cependant être établi à l'avance, pour n'importe quelle entreprise, par un membre du personnel du bureau central de Statistique Canada. Ce cas peut se produire lorsqu'un transporteur a de nombreux LDD, parce qu'un intervieweur dans un LDD donne pourrait ignorer le nombre d'expéditions faites par l'ensemble de l'entreprise. Un autre exemple serait les entreprises qui présentent certaines particularités, comme les transporteurs de marchandises

#### 10. PLAN DE SONDAGE DU DEUXIÈME DEGRÉ D'ÉCHANTILLONNAGE

Enfin, l'échantillon d'entreprises est converti en un échantillon de LDD par l'inclusion dans ce dernier de tous les LDD des transporteurs choisis. représentent une grande proportion des recettes qu'il faut estimer.

qui concerne les principales variables à estimer et que ces transporteurs décision est qu'on sait que cette classe d'entreprises est hétérogène en ce transport routier de marchandises pour compte d'autrui. La raison de cette sélectionnées avec une probabilité de un dans l'enquête de 1981 sur le \$2,700,000 ont été désignées unités take-all c'est-à-dire qu'elles ont été Toutes les entreprises dont les recettes de transport étaient d'au moins

répartition de l'échantillon.

Le premier degré d'échantillonnage consiste à sélectionner dans chaque strate un nombre d'entreprises qui correspond à  $n^4_h$ , le nombre déterminé lors de la

#### 9. PLAN DE SONDAGE DU PREMIER DEGRÉ D'ÉCHANTILLONNAGE



On fait ensuite l'addition de la taille initiale corrigée des échantillons de toutes les strates pour obtenir un total corrigé de la taille initiale de l'échantillon global.

A l'étape suivante, la taille des échantillons est encore modifiée pour que, dans chaque strate, la taille soit égale ou supérieure à celle qu'on aurait obtenue si on avait distribué le total corrigé de la taille initiale de l'échantillon global d'une classe d'entreprises selon le rapport entre la racine carrée du nombre d'entreprises dans chaque strate et le total des racines carrées du nombre d'entreprises dans toutes les strates à l'intérieur de cette classe. Bref:

$$z_h = \max \left\{ \frac{\sqrt{N_h}}{\sum \sqrt{N_h}}, 2^h, 2^h \right\}$$

où la somme est effectuée sur toutes les strates de la même classe que la strate  $h$ .

Enfin, le chargé d'enquête peut subjectivement ajuster la taille des échantillons à  $h$ .

On a retenu cette méthode de répartition de l'échantillon parce qu'il s'agit d'un algorithme qui a donné de bons résultats lors des essais et qui est fondé sur la seule variable qu'on peut mesurer pour toutes les entreprises, c'est-à-dire les recettes faisant partie du champ de l'enquête. Toutefois, il faut rappeler que ces recettes ne sont pas recueillies directement dans l'enquête sur le transport routier de marchandises pour compte d'autrui, mais qu'on transcrit les recettes d'un échantillon d'expéditions. La méthode décrite ci-dessus fait donc abstraction du deuxième degré d'échantillonnage.

D'abord, un programme informatique établit le nombre initial d'entreprises qui doivent être choisies dans chaque strate afin d'atteindre un coefficient de variation cible pour l'estimation des recettes qui font partie du champ de l'enquête à l'intérieur d'une strate. Ce coefficient de variation cible est celui qu'on voudrait obtenir si l'estimation était calculée à partir des recettes totales déclarées par un échantillon d'entreprises choisies, par échantillonnage aléatoire simple, dans une population d'entreprises dont la distribution des recettes comprises dans le champ de l'enquête était la même que celle observée l'année précédente dans l'enquête sur les entreprises de camionnage et de déménagement. La formule est la suivante:

$$N^2 S^2 = \frac{N^2 S^2 + Y^2 (C.V.)^2}{h}$$

où  $N^h$  : le nombre initial d'entreprises qui doivent être choisies parmi les unités non take-all dans la strate h;

$N^h$  : le nombre d'unités non take-all dans la strate h;

$Y^h$  : le total des recettes incluses dans le champ de l'enquête des unités non take-all dans la strate h;

$S^2$  : la variance des recettes incluses dans le champ de l'enquête des unités non take-all dans la strate h;

C.V.<sup>h</sup> : le coefficient de variation cible dans la strate h (la valeur utilisée est la même pour toutes les strates d'une classe donnée mais elle peut varier d'une classe à l'autre).

Deuxièmement, la taille initiale des échantillons est corrigée afin d'assurer qu'un nombre minimal d'entreprises soit tiré de chaque strate, c'est-à-dire:

$$N^h = \min \{ \max (m, N^h), N^h \}$$

où  $N^h$  : le nombre initial corrigé des entreprises qui doivent être choisies parmi les unités non take-all dans la strate h;

m : le nombre minimal d'entreprises à choisir dans la strate h, dans la mesure du possible.

<sup>5</sup> Pour l'année de référence 1981, ce nombre minimal a été fixé à 3 dans toutes les strates.

Une fois que la base de l'enquête est stratifiée, des agents spécialisés peuvent désigner des entreprises comme unités take-all, c'est-à-dire des entreprises qu'ils souhaitent inclure dans l'échantillon avec une probabilité de un. Par la suite, un spécialiste en méthodes d'enquête détermine le nombre d'entreprises à échantillonner parmi les entreprises non take-all dans une strate. Pour obtenir ce nombre, il faut effectuer quelques calculs dont sont exclues les unités take-all.

d'autrui de 1981.

La stratification décrite ci-dessus crée 840 strates, dont 55 n'étaient pas vides dans l'enquête sur le transport routier de marchandises pour compte

surviendront dans la population de l'enquête.

raient varier au cours des prochaines années en fonction des changements qui d'activité (c'est-à-dire \$85, \$350,000 et \$2,700,000) sont souples et pour-

Les valeur limites fixées pour la stratification selon les recettes et le type

Canada. Vingt zones d'exploitation ont été délimitées pour l'enquête.

Nouvelle-Écosse et le Nouveau-Brunswick) mais aucune autre province au des quatre provinces atlantiques (Terre-Neuve, l'Île-du-Prince-Édouard, la "l'Atlantique", ce qui voudrait dire qu'un transporteur dessert au moins deux exploite des services seulement dans cette province. Un autre exemple serait pourrait être le Nouveau-Brunswick, ce qui signifierait qu'une entreprise combinaison spécifique de ces endroits. Par exemple, la région desservie particulier, dans les territoires du Yukon ou du Nord-ouest ou dans une une entreprise assure des services de transport dans une province en comme ceux qui font de grandes expéditions. La région desservie indique si expédition; les autres transporteurs de marchandises générales sont considérés marchandises générales dont les recettes moyennes sont de moins de \$85.00 par les petites expéditions de marchandises générales sont les transporteurs de reuses, de produits agricoles ou d'animaux. Les entreprises spécialisées dans solides réfrigérés, de produits explosifs et (ou) d'autres marchandises dangere-tion, de vrac sec et (ou) de liquides réfrigérés, de machines lourdes, de liquide, de déchargements, de produits forestiers, de matériaux de construc-

Comme la collecte de données coûte très cher à cause des déplacements vers les régions éloignées, on a essayé de restreindre le nombre de LD choisis pour l'échantillon dans la catégorie des transporteurs dont les recettes annuelles dépassent \$350,000. La limite a été fixée à 875 LD par année, soit le nombre utilisé au cours des dix dernières années de l'ancienne enquête.

## 7.2 Nombre total maximum d'expéditions transcrits

La deuxième contrainte d'ordre administratif concerne le nombre total d'expéditions transcrits. Le budget actuel permet de relever au plus 418,000 expéditions, mais ce nombre peut varier d'une année à l'autre selon les résultats des négociations qui ont lieu entre Statistique Canada et les utilisateurs qui participent aussi au financement de l'enquête.

## 7.3 Nombre maximal d'expéditions transcrits par entreprise

Une autre contrainte administrative a été imposée concernant le nombre maximal d'expéditions transcrits pour chaque entreprise. Une limite implicite a été fixée relativement au nombre de jours qu'une équipe de collecte des données peut passer dans un endroit donné, afin que la présence de ces personnes n'in-

dispose pas les répondants.

## 8. STRATIFICATION ET RÉPARTITION DE L'ÉCHANTILLON

A partir des résultats de l'enquête de l'année précédente sur les entreprises de camionnage et de déménagement, les entreprises sont stratifiées en fonction de leurs recettes faisant partie du champ de l'enquête, par type d'activité et par région desservie. Ces variables ont été choisies parce qu'elles illustrent bien le caractère hétérogène de l'industrie. Les recettes faisant partie du champ de l'enquête indiquent si l'entreprise appartient à la classe 1, 2 ou 3, c'est-à-dire si elle a tiré respectivement des recettes de \$2,700,000 ou plus, entre \$350,000 et \$2,700,000 ou entre \$100,000 et \$349,999 au titre du transport interurbain de marchandises au Canada, à l'exclusion des services de déménagement et de transport par véhicule blindé. Le genre d'activité révèle si une entreprise est spécialisée dans les grandes expéditions de marchandises générales, les petites expéditions de marchandises générales, le déménagement sur grande distance, le transport d'automobiles, de pétrole



Lieu lorsqu'une expédition est assurée par un transporteur jusqu'à un point intermédiaire, d'où un autre transporteur continue l'expédition vers un autre point. Les données sur les expéditions de transfert servent à supprimer les doubles comptes d'une expédition.

Les renseignements secondaires qui doivent être consignés sont la date de l'expédition, la quantité de la marchandise et l'unité de mesure (par exemple, 5 pieds-planche, 20 gallons, 15 sacs), des données sur le poids transcrit pour l'expédition (par exemple, le poids minimum, le poids comme mode utilisé pour le calcul des recettes), le tarif exigé et le code des taux (par exemple, un code qui indique si le tarif est minimum, ou exprimé par 100 livres ou par heure) et le code des recettes (c'est-à-dire un code qui indique s'il est impossible d'obtenir les recettes exactes de transport ou si une expédition est exclue du champ de l'enquête).

Les données globales recueillies auprès des petits transporteurs décrivent l'expédition moyenne ou typique en fonction de la province d'origine, de la province de destination, de la marchandise, des recettes moyennes, du poids moyen et du nombre d'expéditions.

## 7. CONTRAINTES ADMINISTRATIVES

La quantité de ressources affectées à la collecte et au traitement des données, de même que l'objectif visant à faciliter la tâche imposée aux répondants ont limité le nombre d'entreprises choisies ainsi que le nombre d'expéditions échantillonnées et de renseignements transcrits.

### 7.1 Nombre maximal d'entreprises dans l'échantillon

La population de l'enquête sur le transport routier de marchandises pour compte d'autrui de 1981<sup>4</sup> est composé de 2,711 entreprises, dont 1,288 ont des recettes annuelles de plus de \$350,000 et 1,423 font entre \$100,000 et \$350,000 annuellement.

<sup>4</sup> L'enquête de 1981 sur le transport routier de marchandises pour compte d'autrui correspond à l'enquête menée en 1982 pour l'année de référence 1981.



Pour cette raison, la base de l'enquête est composée d'une liste de toutes les entreprises dont les recettes annuelles de transport interurbain intérieur dépassent \$100,000. Les entreprises peuvent être aussi divisées en LDD lorsque les documents d'expédition ne sont pas entreposés en un lieu central. La base de l'enquête est établie à partir du recensement annuel des transporteurs pour compte d'autrui effectué par Statistique Canada par le biais de l'enquête sur les entreprises de camionnage et de déménagement<sup>3</sup>.

## 5. UNITÉ D'ÉCHANTILLONNAGE FINALE

L'enquête exploite les données de trois sources, c'est-à-dire celles contenues sur les bandes magnétiques de certains transporteurs, les renseignements transcrits à partir des documents d'expédition échantillonnés et, finalement, des données provenant des transporteurs qui touchent des recettes annuelles de \$100,000 à \$350,000.

Les bandes renferment des données sur les expéditions, dont les caractéristiques sont les mêmes que celles sur les expéditions qui sont échantillonnées et transcrites manuellement. Ainsi, l'unité d'échantillonnage finale est l'expédition, que ce soit pour les entreprises qui fournissent leurs bandes magnétiques ou pour les transporteurs dont les documents d'expédition sont échantillonnés. Dans le cas des entrepreneurs en camionnage qui touchent des recettes entre \$100,000 et \$350,000, on recueille des données globales parce que, habituellement, ces transporteurs ne conservent pas les documents nécessaires sur les expéditions. Pour cette catégorie de transporteur, l'unité d'échantillonnage finale est l'entreprise.

## 6. INFORMATION RECUEILLIE

Les principales caractéristiques qu'il faut noter pour chaque expédition échantillonnée d'un transporteur déclarant des recettes annuelles de plus de \$350,000 au titre du transport interurbain intérieur sont l'origine réelle et la destination finale, la description de la ou des marchandises transportées, le poids et l'unité de poids, les recettes perçues au titre du transport et les renseignements concernant les expéditions de transfert. Un transfert a

<sup>3</sup> L'enquête de Statistique Canada sur les entreprises de camionnage et de déménagement est un recensement annuel des entrepreneurs en transport routier. Elle vise à obtenir des données d'entrée et de sortie sur les établissements, comme les recettes, les dépenses, les renseignements relatifs aux bilans et le matériel exploité.

Les principales contraintes imposées sur le remaniement ont été les suivantes: certains types d'entreprises de camionnage pour compte d'autrui devaient être exclus de la population de l'enquête, notamment les démanagers pour compte propre et les transporteurs de pétrole; la stratification devait être améliorée afin de respecter la structure économique de cette industrie; trois types de données devaient être utilisés à savoir les bandes de certains répondants, les transcriptions provenant d'échantillons de documents d'expédition des LDD d'un échantillon d'entreprises déclarant des recettes annuelles de plus de \$350,000 au titre du transport interurbain intérieur, et les données globales d'un échantillon d'entreprises qui ont touché des recettes annuelles se situant entre \$100,000 et \$350,000 au titre du transport interurbain intérieur; enfin, l'enquête remaniée devait pouvoir être appliquée à l'année de référence 1981, la collecte de données commençant au printemps de 1982.

#### 4. POPULATION ET BASE DE L'ENQUÊTE

La population de l'enquête regroupe toutes les expéditions faites pendant l'année de référence par les entreprises de transport qui satisfont à la définition établie pour le champ de l'enquête. Une expédition est définie comme une quantité de marchandises acheminées par un transporteur et envoyées par une personne ou par un organisme à une autre personne ou à un autre organisme. Les entreprises échantillonnées sont celles dont les recettes annuelles provenant du transport de marchandises entre villes dépassent \$100,000, dont le camionnage constitue l'activité principale et qui sont situées au Canada. Les éléments exclus de cette population sont les expéditions faites par certains types de transporteurs spécialisés comme les transporteurs de pétrole et les déménageurs à compte propre.

Comme cette population idéale n'est pas accessible, on utilise plutôt les entreprises comme grappes naturelles d'expéditions pour l'échantillonnage de premier degré.

des transporteurs. On a donc ajouté une troisième phase à l'étude de l'enquête. Son objectif était d'effectuer les analyses nécessaires afin de formuler et de suggérer des directives générales pour une enquête révisée. Les analyses devaient se conformer aux propositions issues de la phase II.

En juin 1980, le groupe de recherche travaillant sur la troisième phase a proposé que l'enquête soit remaniée de façon à intégrer quatre types de données: les bandes de certains répondants; les transcriptions provenant des échantillons de documents d'expédition tirés de chaque lieu de dépôt des documents (LDD) déclarant des recettes annuelles supérieures à \$1,500,000 au titre du transport interurbain intérieur; les transcriptions provenant des échantillons de documents d'expédition tirés d'un échantillon de LDD ayant des recettes annuelles de \$750,00 à \$1,500,000 au titre du transport interurbain intérieur; et les données globales recueillies auprès d'un échantillon de LDD dont les recettes annuelles au même titre atteignent entre \$100,000 et \$750,000. La décision de recueillir des données globales auprès des petits transporteurs a été fondée sur le fait que ces entreprises ne possèdent pas les documents nécessaires à l'échantillonnage.

### 3.2 Remaniement de l'enquête

#### a) Objectif du remaniement

Après la troisième phase de l'examen de l'enquête, on a décidé de procéder à un remaniement complet de tous les aspects de l'enquête dans le but de recueillir des renseignements plus fiables et plus détaillés sur l'origine et la destination des marchandises acheminées par les transporteurs canadiens pour compte d'autrui. On s'attend à une amélioration non seulement de la fiabilité des données en comparaison avec l'ancienne enquête", mais aussi de la quantité de renseignements recueillis au niveau régional et au niveau des marchandises.

Initialement, l'examen de l'enquête a été décomposé en deux phases.

La première phase avait pour objet d'esquisser des propositions visant à améliorer l'enquête, à l'intérieur du cadre existant, au moyen de seulement quelques ressources supplémentaires. Comme prévu, les propositions ont porté sur une redéfinition de la population de l'enquête, l'amélioration des variables de stratification et l'élargissement de la taille de l'échantillon des documents d'expédition. Les propositions ont été présentées dans un rapport [2].

La deuxième phase avait pour but d'évaluer les propositions formulées, en fonction des besoins des utilisateurs, d'indiquer le coût et les méthodes d'application possibles de chacune des propositions retenues et de poursuivre des analyses de l'enquête. Cette phase a produit une refonte de certaines propositions visant une population d'entreprises plus petite que celle de l'enquête existante et mieux stratifiée en groupes homogènes. En plus de suggérer l'application de ces propositions, on a étudié quatre façons d'agrandir la taille de l'échantillon, à savoir le statu quo, une augmentation de 50 % de la taille de l'échantillon des documents d'expédition, une augmentation de 100 % de cet échantillon et, finalement, une augmentation de 25 % de l'échantillon plus le traitement des bandes de données pour une quarantaine d'entreprises de transport. Après une évaluation des avantages et des coûts de chacune de ces possibilités, la quatrième a été approuvée en principe, parce qu'elle permettait une augmentation considérable de la taille de l'échantillon au coût le plus bas et avec le fardeau de réponse le plus faible. Les recommandations et les détails de l'argumentation figurent dans un rapport [3].

Une première évaluation de la portée des propositions a révélé qu'il fallait effectuer des travaux supplémentaires, surtout pour déterminer le coût total de l'utilisation des bandes contenant les bordereaux d'envoi



### 3. RÉVISION DE L'ENQUÊTE SUR LE TRANSPORT ROUTIER DE MARCHANDISES POUR COMPTE D'AUTRUI

La révision s'est faite en deux étapes principales. En premier lieu, on a procédé à une évaluation détaillée de l'enquête existante. En deuxième lieu, selon les recommandations issues de cet examen, une refonte complète de l'enquête a été effectuée.

#### 3.1 Examen de l'enquête

##### a) Motifs de cet examen

Au début de 1978, Statistique Canada a amorcé un examen de l'enquête sur le transport routier de marchandises pour compte d'autrui pour les raisons suivantes. Premièrement, la Division des transports et des communications de Statistique Canada prévoit toujours une étude périodique des enquêtes en cours. L'enquête sur le transport routier de marchandises pour compte d'autrui n'avait pas été évaluée depuis 1973. Deuxièmement, il était impossible de satisfaire aux besoins courants et prévus de renseignements plus détaillés sur l'origine et la destination des marchandises à l'intérieur des contraintes imposées par l'enquête. Troisièmement, l'expérience acquise par le biais de l'enquête sur le transport routier de marchandises pour compte d'autrui et d'autres enquêtes connexes a fourni des renseignements supplémentaires permettant d'améliorer la base de sondage, les variables de stratification et les techniques d'imputation. Quatrièmement, certaines innovations dans l'industrie du camionnage, comme la production de données sur l'origine et la destination sous forme lisible par une machine, laissaient entrevoir la possibilité d'utiliser des bandes magnétiques pour élargir la base des données et, en même temps, réduire le fardeau des unités déclarantes.

En outre, le caractère de plus en plus complexe des besoins des utilisateurs ont fait sentir la nécessité de perfectionner les techniques de diffusion des données, tandis que des progrès en informatique non seulement ont rendu le système de traitement des données existant désuet du point de vue technique, mais lui ont fait perdre sa rentabilité.



L'enquête sur le transport routier de marchandises pour compte d'autrui

(1969-1979)

C'est en 1969 qu'ont été entrepris les premiers travaux sur une enquête visant à mesurer en fonction de l'origine et de la destination les mouvements de biens entre villes canadiennes effectués par l'ensemble des industries canadiennes de camionnage pour compte d'autrui. A cette époque, une étude a été menée pour examiner diverses méthodes de recueillir les renseignements sur l'origine et la destination des marchandises. Les résultats ont démontré qu'une enquête-échantillon des dossiers administratifs des transporteurs, notamment leurs documents d'expédition, était une méthode convenable de rassembler les données nécessaires.

En 1970, on a procédé à une enquête pilote afin d'évaluer l'efficacité de la méthode de sondage. Pour cette expérience, il fallait examiner les documents d'expédition de 187 entreprises de camionnage pour compte d'autrui dans l'ensemble du pays. La réaction favorable à cette enquête pilote ainsi que la facilité de trouver des renseignements sur l'origine, la destination, les marchandises, le poids et les recettes ont confirmé que la méthode de sondage était applicable.

Ainsi, l'enquête sur le transport routier des marchandises pour compte d'autrui a été menée pour les années de référence 1970 et 1971 avec les objectifs énumérés ci-dessus. Quant à l'année de référence 1972, on a modifié les objectifs de façon à limiter l'enquête aux transporteurs pour compte d'autrui situés au Canada dont les recettes annuelles provenant du transport interurbain étaient d'au moins \$100,000. Pour l'année de référence 1973, on a utilisé une base de sondage mise à jour et mieux définie, qui comprenait les transporteurs routiers réglementés, et on a élaboré une technique d'échantillonnage plus efficace. Depuis l'année de référence 1973, l'enquête se poursuit et les résultats sont publiés annuellement par la Division des transports et des communications de Statistique Canada [1] [5].

La combinaison de tous ces facteurs influe sur le plan d'enquête, surtout pour ce qui a trait à la stratification.

## 2.2 Historique des enquêtes sur l'origine et la destination des marchandises

transportées par des entreprises de camionnage au Canada

### a) Motor Transport Traffic Survey (1957-1963)

La première tentative de mesurer le volume de la circulation des camions au Canada a commencé en 1957 avec l'instauration de la Motor Transport Traffic Survey (MTTS - enquête sur le transport routier), une enquête-échantillon sur les véhicules automobiles utilisés pour transporter des marchandises. La base de l'enquête était une liste de véhicules automobiles immatriculés dressée à partir des dossiers des gouvernements des provinces et des territoires. Cette base était stratifiée selon le type d'activité et le poids brut du véhicule.

La taille de l'échantillon représentait à peu près 10 % de tous les véhicules immatriculés. L'échantillon était prélevé en quatre segments trimestriels dont chacun représentait environ un quart de l'échantillon total. Pour chaque échantillon trimestriel, le travail de collecte des données était réparti en trois semaines d'enquête, un tiers de cet échantillon étant visé pendant une période de sept jours tous les mois.

Comme l'enquête portait sur les véhicules, aucun renseignement n'était recueilli sur l'origine et la destination des marchandises transportées. C'était une enquête sur l'origine et la destination des camions et les marchandises étaient secondaires. On recueillait aussi des données comme la description du véhicule, le nombre de miles parcourus, la quantité de carburant consommé et les coûts d'exploitation du véhicule.

Cette enquête a été menée de 1957 à 1963 inclusivement, puis abandonnée en 1964 par suite de changements dans le système d'immatriculation des véhicules et de la réorganisation de l'industrie du camionnage mais, surtout, parce qu'on avait observé une diminution très importante des taux de réponse.

détriment des autres modes de transport. A la fin des années 1970, Statistique Canada a entrepris, en collaboration avec les principaux utilisateurs, une révision complète de l'enquête.

Le présent document a deux objets: d'abord, présenter quelques renseignements généraux sur l'enquête et décrire les étapes du processus de révision et, deuxièmement, expliquer la méthodologie de l'enquête remaniée qui sera menée pour l'année de référence 1981. Il faut souligner que les détails de la méthodologie de certaines phases n'ont pas encore été mis au point de façon définitive; ce document en décrit toutefois la portée générale.

## 2. RENSEIGNEMENTS GÉNÉRAUX

### 2.1 Bref aperçu de l'enquête sur le camionnage pour compte d'autrui

L'industrie canadienne du camionnage pour compte d'autrui se caractérise par un très grand nombre de petits transporteurs, et la grande diversité des marchandises transportées, des tailles des entreprises et des zones d'exploitation témoigne de son caractère hétérogène.

Les petits transporteurs, c'est-à-dire ceux dont les recettes annuelles sont moins de \$100,000, représentent numériquement 88 % de l'industrie, mais seulement 20 % des recettes d'exploitation. Etant donné l'instabilité de ces transporteurs et leur poids relativement faible dans le total des recettes, il a été décidé de les exclure de la population de l'enquête.

Les entreprises de camionnage assurent le transport d'une grande variété de marchandises, qui exige divers types d'équipement et différents modes d'exploitation. Les transporteurs (déménageurs, transporteurs de marchandises générales, de pétrole en vrac, etc.) se distinguent non seulement par les biens qu'ils acheminent, mais également par le volume des expéditions.

On peut aussi constater le caractère hétérogène de l'industrie du camionnage par l'étendue des régions desservies par les entreprises. Certains transporteurs ne font que des expéditions locales, d'autres des expéditions interprovinciales, tandis que certaines grandes entreprises desservent chaque province ainsi que le marché international.

REMANIEMENT D'UNE ENQUÊTE SUR L'ORIGINE ET LA DESTINATION DES MOUVEMENTS DE MARCHANDISES EFFECTUÉS PAR LES TRANSPORTEURS POUR COMPTE D'AUTRUI AU CANADA<sup>1</sup>

Robert Lussier et Steven Mozes<sup>2</sup>

Ce document présente d'abord quelques renseignements généraux sur l'enquête sur le transport routier de marchandises pour compte d'autrui et décrit les étapes du processus de révision qui a conduit à la décision de remanier l'enquête. Le document explique dans un second temps la portée générale de la méthodologie de l'enquête remaniée qui sera menée pour l'année de référence 1981.

1. INTRODUCTION

L'enquête sur le transport routier de marchandises pour compte d'autrui a été entreprise pour la première fois par Statistique Canada en 1971 afin d'obtenir des renseignements sur l'origine et la destination des mouvements de biens par les soins de transporteurs pour compte d'autrui. Aux fins de cette enquête, l'industrie du transport pour compte d'autrui est définie comme étant l'ensemble des entreprises de camionnage qui, moyennant rétribution, assument le transport de marchandises. L'enquête était fondée sur un échantillonnage probabiliste des expéditions enregistrées sur les documents conservés par les transporteurs canadiens pour compte d'autrui. Or, depuis 1971, la demande de renseignements plus fiables et plus détaillés n'a pas cessé d'augmenter. On peut attribuer cette progression de la demande à de nombreux facteurs, comme la très forte croissance du camionnage depuis le début des années 1950, le caractère de plus en plus complexe des besoins des utilisateurs des statistiques sur le transport, l'intérêt porté à la question de la réglementation de ce secteur économique ou de la levée de normes et, enfin, l'augmentation de la part du marché du transport des marchandises enregistrée par le camionnage au

<sup>1</sup> Ce document est une version révisée d'un exposé présenté sur demande aux Joint Statistical Meetings of the American Statistical Association, the Biometric Society, ENAR and WNAAR, and the Institute of Mathematical Statistics, Cincinnati, 16 au 19 août 1982.

<sup>2</sup> Robert Lussier, division des méthodes d'enquêtes - entreprises, et Steven Mozes, division des transports et des communications, deux divisions de Statistique Canada.





Préparé par Statistique Canada

Comité de rédaction:

G.J.C. Hole  
C. Patrick  
R. Platek  
M.P. Singh  
P.F. Timmons  
H. Lee  
- Rédacteur adjoint  
- Rédacteur en chef

Politique de la rédaction:

La revue "Techniques d'enquête" veut donner aux personnes qui s'intéressent aux aspects pratiques de la conduite d'enquêtes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration de méthodes d'enquêtes: les problèmes de conception causés par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes de recensement et d'enquêtes-ménages, Statistique Canada, 4e étage, Édifice Jean Talon, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à inter-

ligne et demi.



TABLE DES MATIÈRES

Remaniement d'une enquête sur l'origine et la destination des mouvements de marchandises effectués par les transporteurs pour compte d'autrui au Canada	1
ROBERT LUSSIER et STEVEN MOZES.....	
Méthodologie du système canadien d'estimation de l'origine et de la destination des passagers des vols internationaux à horaire fixe	29
GREG HUNTER et LISA DIPLETRO.....	
Certains aspects de la qualité des statistiques sur la mortalité due au cancer et l'incidence de cette maladie	54
D. BINDER et A. MALHOTRA.....	
L'estimation des flux bruts mensuels de l'activité sur le marché du travail	85
STEPHEN E. FIENBERG et ELIZABETH A. STASNY.....	
Remaniement de l'enquête sur le rendement prévu en fruits tendres dans la péninsule de Niagara	111
J. KOVAR.....	
Une estimation précise et rapide de la superficie plantée en pommes de terre grâce à Landsat: Résultats d'une démonstration	129
R.A. RYERSON, J.-L. TAMBAY, R.J. BROWN, L.A. MURPHY et B. MCCLAUGHLIN.....	
Echantillonnage à deux reprises avec pptsr	153
G.H. CHOUDHRY et JACK E. GRAHAM.....	



Canada

Préparé par  
Statistique Canada

numéro 1

volume 9

1983

# TECHNIQUES D'ENQUÊTE





Lacking Vol.9, no.2







JUN 10 1987



